DOCUMENT RESUME

ED 137 317 95 TM 005 909

AUTHOR Burstein, Leigh
TITLE The Use of Data from Groups for Inferences About Individuals in Educational Research. Technical Report No. 7.
INSTITUTION Vasquez Associates Ltd., Milwaukee, Wis.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Dec 75
CONTRACT NIE-C-74-0123
NOTE 190p.

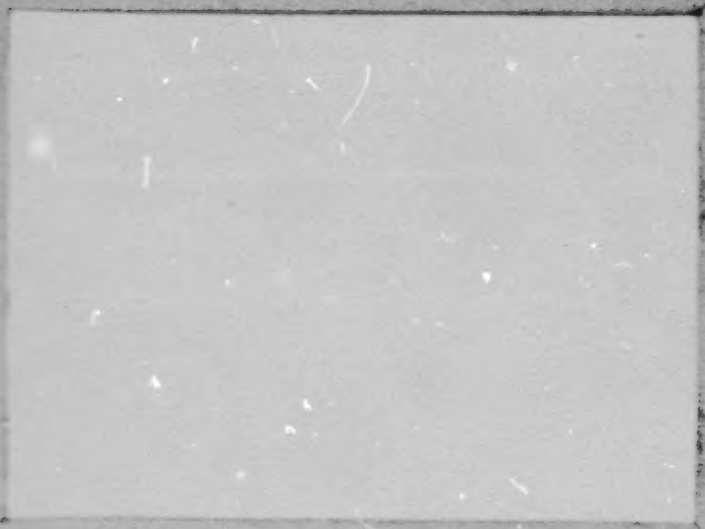
EDRS PRICE MF-$0.83 HC-$10.03 Plus Postage.
DESCRIPTORS *Correlation; *Research Methodology; Sampling; *Statistical Bias
IDENTIFIERS *Grouping (Statistical)

ABSTRACT

Grouping is a statistical procedure through which members of the same group are considered as a single unit of observation. There are methodological and inferential problems associated with various grouping procedures in various settings. This extensive paper focuses on making inferences about individuals when the analysis uses data that is grouped over individuals (for example, school means). The paper identifies five research contexts in which grouping is used, reviews the literature on grouping where only two variables are considered, and proposes a method for clarifying the problems involved in grouping. The method involves introduction of a new variable, a "grouping variable," into the analysis procedure. The grouping variable is essentially the value assigned to members of a group taken over all possible groups. The relationship of the grouping variable to the variables of interest forms the basis for a taxonomy of grouping situations which can then be assessed for certain statistical qualities. The paper discusses the logic and statistical basis for the method, considers the method in a variety of settings, and extends the logic to more complex cases. Empirical examples are presented and considerations for future developments are discussed. (JKS)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED137317



U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM005 909

Vásquez Associates, Ltd.

Consultants for research, evaluation, and management

P. O. Box 5630 Milwaukee, Wisconsin 53211

The Use of Data From Groups for Inferences About Individuals in Educational Research

Leigh Burstein
University of California
Los Angeles

Technical Report No. 7
December, 1975

This research report here was partially supported by National Institute of Education (Contract #NIE-C-74-0123).

Vasquez Associates, Ltd.
1744 N. Farwell Ave.
Milwaukee, Wisconsin 53202

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iii

LIST OF TABLES vi

LIST OF FIGURES viii

CHAPTER 1. Introduction 1

Terminology 1
Inferences Involving Change in the Units of Analysis 2
Research Problems Involving Change in Units 4
Problems to be Considered 13
Overview of Later Chapters 14

CHAPTER 2. Review of the Literature on Grouping Observations . . . 17

Behavioral Scientists' Perspectives on Grouping 18
Econometric Perspectives on Grouping -- "Optimal Grouping" . . 26

CHAPTER 3. Estimation of the Linear Regression Coefficient from Grouped Data in the Single-Regressor Case 29

Terminology and Notation 30
Regression Coefficients to be Contrasted 39
The Bivariate Case -- Standard Model 39
A Structural Model for Determining the Effects of Grouping . . 52
Bias and Efficiency as a Function of Taxonomic Category 67
The Taxonomy as a Guide for Investigation 87

CHAPTER 4. Additional Considerations in the Single-Regressor Case 89

Distributional Factors 90
Scales of Measurement -- Nominal Grouping Characteristics . . . 98

CHAPTER 5. Preliminary Notes on the Multivariate Case 108

Previous Work on the Multivariate Case 108
The "Structural Equations" Approach with Two Regressors 120
The Taxonomy for Two Regressors 126
Implications of Findings 132

CHAPTER 6. Empirical Examples with a Single Regressor 133

Description of Data 134
Regression of Academic Self-Appraisal on Achievement 142
Regression of Achievement on Aptitude 157
Summary of Empirical Results 163

TABLE OF CONTENTS (continued)

Page

CHAPTER 7. Summary and Conclusions 165

Summary of Findings. 165
Suggestions for Further Investigation. 170

REFERENCES. 173

LIST OF TABLES

Table		Page
1.1	Research Problems Involving Data Aggregation.	5
2.1	Correlations of Illiteracy with Race and Illiteracy with Nativity at Different Levels of Aggregation.	22
3.1	Indices of Precision of Estimates from Grouped Data as a Function of Sampling Procedure	38
3.2	Variance-Covariance Matrix for Variables in Equations [3.14a] and [3.14b] (Reduced Forms in Brackets)	56
3.3	Variance-Covariance Matrix for Variables in Equations [3.17a] and [3.17b] (Reduced Forms in Brackets)	58
3.4	Bias θ^* in Estimating Standardized Regression Coefficient β_{YX} from Grouped Data as Function of Group Size, standardized $\beta_{YZ \cdot X}$ and standardized β_{XZ}	76
4.1	Alternative Grouping Variables Based on the Same Grouping Characteristic	92
4.2	Efficiency of Alternative Ways of Grouping on the Same Characteristic as a Function of Sample Size and Number of Groups.	95
5.1	Estimates of Regression Coefficients and Standard Errors with Alternative Grouping Methods from the Houthakker-Haldi Study.	114
5.2	Presence of Bias from Grouping as a Function of Taxonomic Subcategory in the Two-Regressor Case	131
6.1	Questions Included in Composite Self-Appraisal of Academic Abilities (SRAA)	135
6.2	Information on Grouping Variables	137
6.3	Means, Standard Deviations, and Skewness Coefficients of Study Variables, and Zero-Order Correlations of Each Variable with SRAA, ACH, and SAT	138
6.4	Estimates of Parameters Relating ACH(X) and SRAA(Y) to Possible Grouping Variables.	145
6.5	Estimates from Grouped Data of Coefficients Describing the Regression of SRAA on ACH.	149

LIST OF TABLES (continued)

Table		Page
6.6	Comparison of Estimates from Grouped Data Using Different Criteria for Acceptable Bias in the Regression of SRAA on ACH.	154
6.7	Weighted Composites from Grouped Estimates of β_{YX} from the Regression of SRAA on ACH	156
6.8	Estimates of Parameters Relating SAT(X) and ACH(Y) to Alternative Grouping Variables (Z).	159
6.9	Estimates from Grouped Data of Coefficients Describing the Regression of ACH on SAT.	161
6.10	Comparison of Estimates from Grouped Data Using Different Criteria for Acceptable Bias in the Regression of ACH on SAT	162

LIST OF FIGURES

Figure		Page
3.1	Path Diagrams Corresponding to Categories of the Taxonomy.	66
3.2	Aggregation Bias θ^* (as defined by [3.31]) as a Function of Standardized β_{XZ} and Group Size n (with $\beta_{YZ \cdot X}$ fixed at .1)	74
4.1	Path Diagram Incorporating Both Latent and Manifest Grouping Variables	102
4.2	Path Diagram for Aggregate Data Grouped by Z^+	103
5.1	Path Diagrams for the Subcategories of the Taxonomy in the Two-Regressor Case.	128-130

CHAPTER 1

INTRODUCTION

Problems of data aggregation have important implications for educational research utilizing data from groups of individuals. This investigation considers the consequences of "change in the units of analysis" where relations among individuals are to be inferred from grouped data. In Chapter 1 five research problems are discussed where an investigator might attempt to translate properties and relations from one level of grouping to another. A general strategy is described for analyzing the conditions under which grouped data can be used for inferences about individuals.

I. Terminology

Data aggregation denotes the replacement of a set of numbers by a smaller set of numbers or "aggregates". This term occurs repeatedly in the literature of economics and econometrics; macroeconomic theory is based mainly on aggregated measurements.

Whenever distinct measures are combined, aggregation is involved. In a study of foods, products such as bananas and oranges can be combined into the category "fruits". A single aggregate index such as per capita consumption of all fruits can replace separate consumption indices for bananas and oranges. In an educational context, the mean aptitude score attributed to a school is an aggregate of the scores of its students.

Measurements can be aggregated within a person as well as between persons. In observational studies, the observation period is usually

divided into time intervals. Behavior during a given time interval is represented by either the total or average number of occurrences of the behavior in that period. Such a score is an aggregate of instantaneous observations.

Here, grouping of observations or, simply, grouping, will refer to the aggregation of measurements over individuals (as distinct from aggregation over time periods or commodities). More specifically, grouping is the replacement of numbers representing observations on individuals with a smaller set of numbers representing observations aggregated over a group of the individuals. An example is the formation of school means from the achievement of students. The terms aggregation, grouping, and grouping of observations will henceforth be used interchangeably.

II. Inferences Involving Change in the Units of Analysis

Grouped data are common in the social sciences. Sociologists focus on relations among collections of individuals. Educational researchers often use the classroom or school as the sampling unit and analyze between-class and between-school relations. The study of grouped data introduces no special obstacles when inference is restricted to the level at which the data are analyzed. If a study concerns the relation between the academic and social psychological climates of the school, the school-aggregated achievement and attitude indices are the data to relate.

On the other hand, educational and psychological researchers are usually concerned with relations among measurements on individuals. The investigator may wish to determine the relation between student aptitude and student achievement or between parents' education and

student aspirations. These measurements cannot always be examined at the individual level, possibly because those data are not obtainable or identifiable for each person, or perhaps because of high cost of analysis.

Facing such problems, some investigators turn to data on groups to estimate regression and correlation coefficients at the individual level. Their conclusions extend the results of the analyses at the group level to the relations among individuals.

However, complications arise in translating properties and relations from one level to another (Riley, 1964; Robinson, 1950; Scheuch, 1966; Theil, 1954; Thorndike, 1934). These complications are discussed under the general label "change in the units of analysis". Where this problem arises, the investigator wishes to apply relations observed among units at one level of aggregation to units at a different level (Blalock, 1964). The direction of inference can go toward larger aggregates, such as states or nations, or toward smaller ones -- the smallest being the single person.

Our concern is with research where relations at the individual level are of interest, but data are aggregated over individuals according to some specifiable grouping rule.[1] The criterion for grouping can be almost any characteristic of the individuals. Grouping can even be random. The choice of grouping procedure is dictated by the data on hand and the usefulness of a specific procedure for estimating individual-level relations from these data.

---

[1]Terms such as "grouping procedure", "grouping method", "grouping rule", "grouping technique", and "grouping strategy" will be used interchangeably in referring to the formation of groups. "Grouping characteristic" and "grouping criterion" will refer to the characteristic(s) from which the group classifications are determined. The actual classification scheme which assigns observations to groups will be called the "grouping variable".

III. Research Problems Involving Change in Units

We next describe five research problems in which grouped observations are used in estimating relations among measurements on individuals. Missing observations, fallibly measured variables, economy of analysis, anonymously collected information, and ecological inference all create problems that data aggregation can alleviate to some degree.

The degree of investigator control over the aggregation of data is an important consideration. In certain contexts group membership is determined in some natural way, e.g., school attended or census tract. It is thus beyond the investigator's control except for the exclusion of sampling units and individuals (limited or no investigator control). In other contexts the investigator can manipulate the formation of groups (completely or partially). There are generally more options for improving estimation in the latter contexts. In Table 1.1 we indicate the degree of investigator control over the formation of groups for each problem. Why the methods of data aggregation are used, how such methods are applied, and where they are principally applied are also discussed.

A. Missing Observations

Missing data are to be expected whenever an investigator collects a large amount of information or uses a large number of subjects. Missing observations are particularly likely in longitudinal studies. Thus, if student achievement is assessed on three occasions, a particular student may miss one or more testing periods. The investigator must then decide how to treat this hiatus in estimating the relations among the tests, or the relations of the tests to other

Table 1.1. Research problems involving data aggregation.

Research Context	Reasons for Data Aggregation	Description of Application	Principal Application
I. Complete Investigator Control -- Group membership can be defined by any characteristic in the data set which is measured for all individuals.			
(A) MISSING OBSERVATIONS	Missing observations on primary variables for some individuals inhibit confidence in analytical results.	Each missing observation on a primary variable is replaced by the mean response on that variable in some group to which the individual belongs.	Longtitudinal and cross-sectional analysis of survey data.
(B) FALLIBLY MEASURED VARIABLES	Random errors of measurement associated with independent variables attenuate regression coefficients.	Different approaches have been suggested as part of the general refinement of statistical procedures for handling "errors-in-variables" problems.	Statistical treatment of measurement errors.
(C) ECONOMY OF ANALYSIS	Budgetary constraints make analysis of massive data bases at the individual level impractical.	Data are collapsed into a smaller number of units by some grouping rule.	Analysis of census data and national, regional, and state school statistics.
II. Partial Investigator Control -- Group membership can be defined by any characteristic which has been measured simultaneously with each primary variable.			
(D) ANONYMOUSLY COLLECTED INFORMATION	Data on certain primary variables are collected anonymously, making it impossible to match observations on primary variables at the individual level.	Characteristics measured simultaneously with the anonymously collected information can be used to aggregate the data.	Confidential student records and responses to attitudinal questionnaires.

Table 1.1. (continued). Research problems involving data aggregation.

Research Context	Reasons for Data Aggregation	Description of Application	Principal Application
III. Limited or No Investigator Control -- Group membership is determined prior to the collection and analysis of data; group membership is directly pertinent to the study of primary variables.			
(E) ECOLOGICAL INFERENCE	The sampling units of the investigation constitute "natural" aggregates of individuals.	Disaggregation efforts are generally a necessary precondition to reasonable inferences at the individual level.	Analysis of school and classroom means where the school and the class are the sampling units; data organized by census tract or demographic region.

variables (school characteristics, teacher characteristics, home environment, etc.).

Many investigators simply drop from the data set any individual who lacks information on any study variable. However, Elashoff and Elashoff (1971, p. 1) find that "techniques such as case deletion which assume that observations are missing at random may be extremely misleading. If the probability model governing the occurrence of missing data is complex, the only adequate solution may be to find out what the missing observations are".

Some investigators use the mean of the overall sample or the mean of some subgroup to which the individual belongs as an estimate of the missing observations. This "replace-with-the-mean" strategy is somewhat akin to the adjustments made in factorial analysis of variance (ANOVA) experiments where a missing observation (X_{ijk}) is estimated by the mean of the ijth cell ($\bar{X}_{ij.}$).

The replace-with-the-mean strategy is a use of aggregated data. For example, Kline, Kent, and Davis (1971), investigating the political instability of nations, replaced missing observations on stability and literacy with means. These means were estimated from subgroups of nations grouped by variables measured on all nations (date of independence, location, political modernization). So each nation with missing data on stability is assigned the mean stability score estimated from its subgroup on, say, date of independence.

The utility of replacing missing observations with group means depends on the variables under study. The estimates generated are functions of the properties of the grouping characteristics -- their internal distributional properties and their relations with the study

variables. A good estimate of the actual relations can be obtained because information relevant to the problem has been retained and information loss thereby reduced. On the other hand, as with case deletion, certain questions remain. In fact, the treatment of the missing data can be complicated as well as simplified by this particular grouping strategy.

B. Fallibly Measured Variables

It is well known that estimates of regression coefficients are attenuated by random errors in the independent variables. Let β_{YX} be the regression coefficient where X is the observed independent variable and let $R_{XX'}$ represent the reliability of the measurement of X_∞, the person's true score on the independent variable. The usual procedure is to use $\beta_{YX}/R_{XX'}$ rather than β_{YX} to estimate β_{YX_∞}.

Madansky (1959) reviewed in detail the literature on the fitting of straight lines when both variables are subject to error. He discussed several grouping techniques that were proposed to handle problems arising from an imperfectly measured independent variable in regression analysis. Methods developed by Wald (1940) and Bartlett (1942) are perhaps the most familiar.

Recently, Blalock (1964; 1970) has reconsidered the Wald-Bartlett techniques and has advanced his own plan for using grouping with imperfectly measured variables. He recommends that the investigator group on an "instrument", a variable which (1) affects the "true" independent variable, and (2) does not directly affect the dependent variable. The relationships of interest are then estimated from the grouped observations.

Both the Wald-Bartlett and Blalock grouping techniques are based on the principle that measurement errors tend to cancel out within

groups if the grouping characteristic is highly related to the "true" values of the independent variable but is uncorrelated with the measurement errors. Under these conditions, the error portion of the observed variance in the independent variable decreases when group means are used, especially as the size of the groups becomes large. Thus the reliable portion of the variation increases through grouping, and the regression estimates are in part disattenuated.

The merits of the approaches suggested by Wald, Bartlett, Blalock, and others will not be debated here. However, their work suggests that the grouping of observations may be one way to resolve certain measurement difficulties.

C. The Economy of Analysis

Grouping may be prescribed when there is an overabundance of relevant data, and the budget for analysis is limited. For example, costs may prevent use of the complete data from the California State Testing Program in relating minority status to achievement. The analyst may choose to sample districts, or to carry out a between-districts analysis. The latter analysis involves a change of units if the investigator then makes interpretations at the individual level.

Econometricians have already developed sound principles for grouping where economy of analysis is the chief concern (Prais and Aitchinson, 1954; Cramer, 1964). The resulting loss of efficiency has been only a few percent in the cases economists typically examine.

The successful use of aggregation in this context can be largely attributed to the investigator's ability to choose the aggregating variable. In most cases where economy of analysis is the concern, the investigator can choose which characteristic(s) will define the groups

of students, be it sex, classroom, school, or some other measure. The investigator can secure meaningful estimates from aggregated data by choosing a grouping characteristic whose relations with the study variables best satisfy the conditions of efficient grouping.

D. Anonymously Collected Information

It is usually impossible to match data collected anonymously with identified information on other variables on the same persons. For example, student achievement cannot be compared with attitude when responses to the attitude questionnaires are anonymous. If, however, some auxiliary information about individuals can be collected along with the attitude questionnaire, partial identification by group membership can sometimes lead to accurate and efficient estimation of the relations between attitudes and achievement (Boruch, 1971; Feige and Watts, 1970; 1972). These estimates may be obtained from grouping procedures similar to those used in contexts where the investigator has complete control over the choice of grouping procedure. For example, students can be grouped by county of residence; then the between-county regression of student attitude on student achievement can be used as an estimate of the individual-level regression. Or the student could be asked to indicate the second letter of his last name. What auxiliary information is suitable depends on the study conditions, but a "good" grouping technique has certain general properties. Once these properties are known, the investigator can build "good" grouping characteristics into his study design.

E. Ecological Inference -- Aggregate Sampling Units

It is not uncommon to sample aggregates of individuals rather than the individuals themselves. For example, every student in a classroom

can be studied rather than students selected individually from the student body. Scores can be obtained from student bodies of schools and colleges, and between-school and between-college relations analyzed. City, county, and census tract means can be the sampling units in sociological and economic studies.

Inferences drawn from aggregate sampling units can lead to what has been called the "ecological fallacy" (Robinson, 1950). The "ecological fallacy" is the practice of inferring relations between properties of individuals from the relations of group data (Alker, 1969; Selvin, 1958). Although sociologists and political scientists beginning with Robinson have discussed "ecological inference", the writers in the educational and psychological literature have often overlooked the issue.[2]

When sampling units are groups of individuals, between-group analysis is logical even when the relations among measurements on individuals are the primary concern. The investigator lacks control over group membership and thus cannot select a suitable grouping procedure as in other contexts. In many instances, he is unable to determine how the required grouping procedure affects the variation and covariation of the study variables. Under these conditions, the possibility of inferring relations at the individual level is limited.

In any case, the sampling of groups can present a particularly complex type of aggregation problem, since questions regarding sampling

---

[2]Oddly enough, one of the first references to the inflationary effects of estimating correlation coefficients from grouped data was by the eminent psychologist E.L. Thorndike (1934). There appear to be no further comments on the topic from educational and psychological researchers except the papers questioning the appropriateness of estimating individual learning curves from grouping learning curves. (e.g., Estes, 1956).

bias arise in addition to concerns about level of inference. One question may be whether the sampled classrooms (counties) are representative of the classrooms (counties) in the universe to which one wants to generalize. The investigator must clearly understand the basis for his inferences to the individual level in order to be at all confident. Otherwise, it may be best to make inferences at the group level or to examine the individuals within groups, or to do both.

F. Applicability of Grouping Scheme in Different Contexts

This investigation offers a general scheme for identifying the consequences of grouping. This scheme will enable an investigator to choose the best grouping characteristics from a larger set when information about interrelations of each grouping characteristic and the study variables is known. Thus, our results are most applicable to problems (A) through (D) where the investigator has at least partial control over the aggregation procedure. The ordered grouping characteristics that can occur in these contexts are also easier to handle since the determination of the relations of ordered characteristics to the study variables is straightforward.

The extra difficulties of grouping when some data are collected anonymously [problem (D)] largely arise from the inability to group on certain primary variables. It is best to group on the independent variable in a regression analysis, but this is not possible when observations on independent and dependent variables cannot be linked. The general scheme will offer suitable alternative procedures in this context that approximate the optimal grouping method.

The problems of ecological inference [problem (E)] are the most complex because there is no choice of grouping procedure and also because the observable grouping characteristic usually has a nominal

scale. Our scheme offers little direct guidance on how to proceed in this context, though it will usually indicate when inferences about individual relations are out of the question. However, the conditions necessary to determine when such inferences are reasonable are unlikely to occur unless the analysis can be carried out at the individual level. If individual-level analyses are possible, ecological inference is usually unnecessary.

The analytical arguments will be restricted mainly to the conditions prevailing when the investigator can choose among several ordered grouping characteristics [problems (A) through (D)]. Our hypothetical examples and empirical analyses will refer mainly to problems of economy of analysis [problem (C)] and of anonymously collected information [problem (D)]. Application in other contexts will be indicated where appropriate.

IV. Problems to be Considered

This inquiry focuses on how grouped data can be used for inferences about individuals particularly in educational research. The problems discussed in the previous section affirm the need for a clear understanding of this technique. We cannot specify the problems exactly until the technical terminology and notation are developed, but we can identify previously unsettled issues to be considered.

Regression and correlation coefficients calculated from grouped data may be biased estimates of the corresponding individual-level relations. Robinson (1950) showed that the bias in such correlation coefficients is strongly determined by the ratios of the between-group variation of the variables to their total variation. Other researchers (Blalock, 1964; Feige and Watts, 1972; Hannan, 1970; 1971) have shown

empirically that the bias in a regression coefficient depends on the relation of the grouping characteristic to the independent and dependent variables.

We propose to trace rationally how aggregation bias depends on the interrelations among the variables of interest. Our structure, which includes cases hitherto neglected, will be a taxonomy that contains the possible interrelations between the grouping variables and the other variables. In addition to presenting logical and empirical arguments, as in previous studies, we shall develop mathematical formalization for the effects due to the choice of grouping variable. While emphasizing bias, we shall also discuss efficiency and precision of regression coefficients. Bias in correlation coefficients will be considered only incidentally although a way of estimating zero-order correlations from grouped data is also described.

Aggregation bias will be studied in both bivariate and multivariate relationships. The effects of varying the number of groups and the number of observations per group will also be considered. The latter work will indicate which properties of grouping are most sensitive to sample size. The intent is to formulate strategies for minimizing information loss when grouped data are used for individual inference.

V. Overview of Later Chapters

Earlier literature on estimating correlation and regression coefficients from grouped observations is reviewed in Chapter 2. Most of the work cited is drawn from sociology and economics.

In Chapter 3 we state formally what is known about estimating the simple linear regression coefficient from grouped observations and extend previous work. Alternative models are discussed. After

extending the "structural equations" approach (Blalock, 1964; Hannan, 1970, 1971; 1972) by incorporating a function of the grouping characteristic as a variable in the system, we present a taxonomy of the relations between the "grouping variable" and the other study variables. Formulas are then derived for the bias and efficiency of variables from each taxonomic category. Finally, we discuss the implications of the results for investigators using grouped data.

Other aspects of the single-regressor case are considered in Chapter 4 with emphasis on within-variable factors such as the number of groups and the number of cases per group. We also describe an alternative scheme for characterizing the grouping process which complements the treatment in Chapter 3. The chapter closes with a discussion of ways to examine the effects of grouping on a nominal characteristic.

In Chapter 5 we consider the case of two regressors and point toward extension to any number of additional independent variables. The literature specific to the multivariate case is reviewed, and the taxonomic approach is applied to the two-regressor model.

An empirical demonstration of effects in the single regressor case is presented in Chapter 6. Information collected from incoming freshmen at a large Midwestern university serves as the data base. First, for a certain X,Y pair, the regression slope and its sample variance are estimated from the ungrouped observations under a simple linear model. Then one or another student characteristic is used to group observations, and the Y-on-X regression slope and its sample variance are estimated from the grouped observations. The empirical results are shown to conform to the conclusions derived analytically.

The use of composites of estimates from different grouping procedures is described; this improves estimation in certain contexts.

Chapter 2

REVIEW OF THE LITERATURE ON GROUPING OF OBSERVATIONS

Historically, investigations of the effects of grouping on the estimation of individual-level relations have followed two distinct lines of inquiry. On the one hand, statisticians and behavioral scientists (mostly sociologists) have considered this question while studying the "ecological fallacy" (Robinson, 1950), the effects of modifiable units (Yule and Kendall, 1950), and the problems caused by a "change in the units of analysis" (Blalock, 1964). These investigations share an interest in the circumstances under which the analysis of grouped units inflates estimates of individual-level relations.

Economists, on the other hand, have traditionally treated grouping as a legitimate strategy for reducing the cost of analysis. Their mathematical formulations have indicated that grouping simply reduces the efficiency of regression estimates without introducing any bias. Thus, they have hunted for the most efficient means of forming groups. Prais and Aitchinson (1954) and Cramer (1964) represent this econometric tradition.

In recent years, the distinctions between the approaches have blurred as the methodologies of the behaviorial sciences and econometrics converged. Hannan (1970, 1971; 1972) and Feige and Watts (1972) are largely responsible for this convergence.[1]

Below, we review only the key presentations from the two lines of inquiry. Summaries of previous work in these areas have already appeared

---

[1]See also Burstein (1974) and Hannan and Burstein (1974).

elsewhere.[2] We reserve the detailed discussions of certain key investigations for a later chapter. In Chapter 3 we examine work by Prais and Aitchinson (1954), Cramer (1964), Blalock (1964), and Hannan (1971; 1972) on the effects of grouping on the estimation of simple linear regression coefficients. In Chapter 6 our study of the multivariate case is juxtaposed with reviews of work by Prais and Aitchinson (1954), Haitovsky (1966), and Feige and Watts (1972).

I. Behavioral Scientists' Perspectives on Grouping

The earliest articles on the effects of grouping indicated that correlation coefficients increase when the size of units (e.g., census tracts) is increased. In 1934, Gehkle and Biehl showed how the correlation of total number of male juvenile delinquents with median monthly rental in Cleveland, Ohio changed from -.502 as the city's 252 census tracts were successively grouped into larger regions. The magnitude of the correlations increased steadily with the degree of aggregation:

	NUMBER OF REGIONS							
	252	200	175	150	125	100	50	25
CORRELATION	-.502	-.569	-.580	-.606	-.662	-.667	-.685	-.763

[2] Selvin (1958), Scheuch (1966), Alker (1969), Allardt (1969), Cartwright (1969), Shively (1969), and Iversen (1973) among others reviewed the grouping literature in the behavioral sciences, focusing on Robinson's (1950) work and related papers but offered little significant new material. Among the above, only Selvin and Scheuch refer to related studies by Gehlke and Biehl (1934), Thorndike (1939), and Yule and Kendall (1950). Johnston (1971) reviews the econometric studies. Hannan's work (1970, 1971; 1972) combines a review of previous work with contributions to the theory.

Thorndike (1939) demonstrated the problems associated with the use of grouped data in the course of his investigation of the determinants of intelligence. He pointed out that the correlation between two traits (X and Y) in m groups equals the correlation between the traits for the individuals composing the groups only under very special circumstances. He added that the latter correlation was usually much smaller.

Thorndike then constructed an illustration with intelligence quotient as X , and the number of rooms per person as Y , and the twelve districts of a city as units for aggregation. Within each district he created a sample of X and Y values such that within districts $r_{XY} = 0$. When observations at the individual level were subsequently pooled over districts, $r_{XY} = .45$; but the between-districts correlation of X and Y averages was .90.

More than ten years passed before questions regarding inferences from grouped data reappeared. Yule and Kendall (1950) stated that if the units of analysis were modifiable (e.g., characteristics of geographical regions), the magnitude of a correlation depended on the unit chosen. Accordingly, correlations "measure the relationship between the variates for the specified units chosen for the work" (Yule and Kendall, 1950, p. 312). Furthermore, they concluded that whenever units are grouped and correlations are calculated from summary characteristics of the groups, such as averages, the correlations increase with the size of the grouping. Conversely, coefficients decrease as the grouping becomes finer. As we shall see, this generalization is now known to be incorrect.

In addition to their citation of the Gehkle and Biehl example, Yule and Kendall correlated the yields of wheat and potatoes from 48 agricultural counties in England in 1936 and successively halved the

number of units by combining contiguous areas (forming 24, 12, 6, and 3 units). These groupings yielded correlations of .219, .296, .576, .765, and .990, respectively.

Sociologists and political scientists dominated the literature dealing with grouping for most of the next twenty years. The early sociological investigations typically focused upon bivariate relations between qualitative variables where the observations were grouped by location (e.g., state), by social organization (e.g., school), or by temporal occurrence (e.g., quarterly statistics). Investigators were generally concerned about the consequences of using such data to make inferences about the ungrouped observations. These analysts' problems were amplified by their lack of control over the grouping process.

The article by Robinson (1950) on the "ecological fallacy" "... triggered one of the liveliest methodological debates in the postwar period" (Scheuch, 1966, p. 148). Alker (1969) described the surprise, dismay, and rage of users of ecological data that Robinson caused with his demonstration that statistical associations for aggregated populations can differ in magnitude and even in sign from those for individuals. Robinson advised a distinction between "individual correlations", which he defined as a correlation between indivisible objects, and "ecological correlations", where the statistical objects are defined as a group of persons. He warned against treating ecological correlations as if they were individual correlations. Robinson considered it to be an "ecological fallacy" to use data grouped by territorial units as if they were measurements on individuals.

The avowed purpose of Robinson's paper was to provide a mathematical formulation of the exact relation between ecological and individual correlations and to show how that relation reflected upon the practice

of using ecological correlations as substitutes for individual correlations.[3] His analyses on race vs. illiteracy and race vs. nativity (see Table 2.1 below) are illustrative.

Robinson's explanation can be summarized as follows:

i) The individual correlation depends upon the internal (within-cell) frequencies of the within-areas contingency tables, while the ecological correlation depends upon the marginal frequencies of the within-areas contingency tables.

ii) Since the within-group marginal frequencies from which the ecological correlation is computed do not fix the internal frequencies, which determine the individual correlation, there need not be any correspondence between the individual and ecological correlations.

According to Robinson, the mathematical relation between individual and ecological correlations can be written as

$$[2.1] \qquad r_E = k_1 r - k_2 r_W \quad ,$$

where

$$k_1 = 1/\eta_{XY}$$

and

$$k_2 = \sqrt{1 - \eta_X^2}\,\sqrt{1 - \eta_Y^2}/\eta_X \eta_Y \quad .$$

In these equations, r is the correlation between X and Y for all N persons; r_E is the "ecological" correlation, the weighted correlation between m pairs of X and Y percentages which describe

---

[3] In Robinson's opinion, ecological correlations were used simply because measures on individuals were not available. Others, beginning with Menzel (1950), pointed out that relations among the properties of collectives can have their own inherent value. Questions regarding appropriate units of analysis remain outside the domain of this investigation. We are only interested in inferences about the relations at the level of individuals when the analysis is performed on grouped data.

Table 2.1. Correlations of illiteracy with race and illiteracy with nativity at different levels of aggregation[a].

Description of Units	Value of r (illiteracy and race)	Value of r (illiteracy and nativity)
97,272,000 persons	.203	.118
48 states	.773	-.526
9 geographic regions	.946	-.619

[a]The correlations are Pearsonian fourfold correlations based on data from the 1930 U.S. Census. The three attributes are all dichotomous (literate vs. illiterate; Negro vs. Non-Negro; Native-born vs. Foreign-born).

the subgroups in a fourfold table; and r_W is the average of the m within-group correlations between X and Y, each within-group correlation being weighted by group size. Also, η_X^2 and η_Y^2 are the correlation ratios (the ratio of the between-group variation to the total variation) which measure the degree to which values of X and Y cluster by group.

From equation [2.1], Robinson was able to deduce that the individual and ecological correlations are equal only when

$$[2.2] \qquad r_W = k_3 r \quad ,$$

where

$$k_3 = \frac{1 - \eta_X \eta_Y}{\sqrt{1 - \eta_X^2}\sqrt{1 - \eta_Y^2}} \quad .$$

However, since the minimum value of k_3 is unity,[6] the individual and ecological correlations can be equal only if the average within-group correlation is larger than the individual correlation. This is counter to experience; hence there is no reason to expect equivalence of the ecological and individual correlations.

---

[6]In the unlikely case that either correlation ratio equals 1, the value of k is undefined. Otherwise, for any two numbers a and b,

$$1 \leq \frac{1 - ab}{\sqrt{(1-a^2)(1-b^2)}}$$

$$1 - a^2 - b^2 + a^2b^2 \leq 1 - 2ab + a^2b^2 \quad \text{(multiplying by the denominator and squaring both sides)}$$

$$0 \leq a^2 + b^2 - 2ab$$

$$0 \leq (a-b)^2$$

and thus, since we can let $a = \eta_X$ and $b = \eta_Y$, the minimum value of k_3 is 1.

Equation [2.1] also suggested to Robinson how the ecological correlation depended upon the number of subgroups. He pointed out the following effects of consolidating units:

i) The ecological correlation decreases as the groups become more heterogenous since r_W increases directly with increasing group size and the between-group proportion of the variation equals $1 - r_W^2$.

ii) The correlation ratios η_X^2 and η_Y^2 decrease as the between-groups variation becomes smaller.

iii) Of the two effects, the changes in the correlation ratios are considerably more important than the changes in r_W so that the numerical value of the ecological correlation increases with increasing consolidation of units.

After Robinson, the emphasis in studies of the effects of grouping shifted to a search for conditions under which the bias from grouping can be minimized. Duncan and Davis (1953) developed an estimate of the size of the error when aggregated data are used to predict individual-level relations. They examined successive subdivisions of a territorial unit (in their example, census tracts) and used the differences in the ecological correlations that were obtained for the units of varying size as the best estimate of the size of the ecological fallacy. They concluded that "although different systems of territorial subdivision give different results, ... the criterion for choice among these results is clear. The individual correlation is approximated most closely by the least maximum and the greatest minimum amongst the results from several systems of territorial subdivision" (Duncan and Davis, 1953, p. 666).

Goodman (1953; 1959) proposed the use of ecological regression coefficients, rather than ecological correlations, in any attempt to

define the circumstances that reduce the problems Robinson had identified. Goodman's form of ecological regression is appropriate for variables which are measured nominally or ordinally, and his method though requiring some difficult assumptions, is more efficient than the Duncan-Davis method of setting bounds.

Briefly, his method is as follows. Let Y be the proportion of the total population who are illiterate, X be the proportion of the total population who are Negroes, p be the proportion of Negroes who are also illiterate, and q be the proportion of Whites who are also illiterate. Finally, let the groups represent samples from the population of X and Y values. Then, if (a) population parameters p and q do not differ from area to area and (b) $E(Y) = Xp + (1 - X)q$ -- where X is as defined above and $E(Y)$ is the expected proportion of illiterate people in an area -- the standard least-squares approach yields unbiased estimates of p and q and thereby of the slope of the regression of Y on X. Furthermore, if the values of Y are approximately normally distributed with the same variance for each value of X, all standard regression methods also apply.

Thus, according to Goodman (1959, p. 614), the only assumption necessary to justify his estimation procedures is that p and q "must be more or less constant for the different ecological areas in such a way that the standard linear regression model can be applied". His estimates of the individual-level parameters in the Robinson and Duncan-Davis examples were a vast improvement over those from ecological correlations or the Duncan-Davis bounds.

Blalock's examination of "change in the units of analysis" problems was the first break from the consideration of exclusively nominal and ordinal variables. Blalock (1964) used a causal framework to examine

empirically the effects of grouping strategies on the correlation coefficient r_{XY} and the regression coefficients b_{YX} and b_{XY}. He placed artificial restrictions on the grouping criterion in order to alter the variation among X and Y variables in specific ways: to maximize variation in X, to maximize variation in Y, and to minimize the effects of grouping on both variables (random grouping). Fourthly, areal units were grouped by proximity.

Blalock demonstrated that r_{XY} remained unchanged only under random grouping. When Y was the dependent variable, both random grouping and maximizing variation in X left the estimate of b_{YX} unchanged; but the variance of the slope estimate increased. However, b_{XY} was affected by maximizing variation in X. Thus, if one is to infer individual-level relationships from aggregated data, individuals have to be grouped in such a way that their scores on the dependent variable are related to group membership only indirectly, through their scores on the independent variable.

II. Econometric Perspectives on Grouping -- "Optimal Grouping"

Econometricians have traditionally followed an entirely separate line of inquiry. The problems they have attempted to solve are those caused by an overabundance of data. They consider the practical problems facing an investigator who can choose among a variety of grouping methods. Prais and Aitchinson (1954) and Cramer (1964) have done basic work to be recounted in detail later. Here, we provide only a short summary.

Within a general regression model, Prais and Aitchinson (1954) set out to estimate the regression parameters $\beta_{YX_1}, \ldots, \beta_{YX_K}$ for K regressors, and the variances of the estimators from the individual and

grouped observations. Following classical least-squares procedures, they showed that, whatever the method of grouping (a) the resulting estimators are always unbiased, (b) the variances of the estimators based on grouped data are always greater than those of the estimators from the original observations, and (c) the efficiency of grouped estimators is optimized by maximizing the between-groups variation in the regressors.

For most of the 1950's and 60's, the Prais-Aitchinson results defined the state of the econometric knowledge on the topic. Cramer (1964), following the Prais-Aitchinson approach, focused on strategies for optimal grouping in the two-variable case without seriously considering the possibility of bias. He evaluated certain efficient grouping procedures under conditions common to economic survey analysis and provided empirical examples on optimal grouping from the literature on economics.

Haitovsky (1966; 1973) did not follow the path laid out by Prais-Aitchinson and Cramer. Instead, he studied alternative ways of estimating multiple-regression coefficients when the data are in the form of one-way classification tables for which the cell frequencies of the cross-classifications are not available. His most important contribution is his empirical evidence that grouping on one regressor can lead to biased estimators when the hypothesized model contains multiple regressors.

Recent work by Feige and Watts (1972) is even more definitive in the multiple-regressor case. They considered the analytical consequences of "partial aggregation" as a means of performing individual-level analysis while preserving the confidentiality of data. Perhaps this new substantive focus explains how they found differences between estimators

of regression coefficients based on the individual and grouped data, a result contrary to the findings of Prais and Aitchinson but in accordance with Blalock's findings. They attributed the differences to one of three sources: (i) specification bias (omission of regressors), (ii) bias introduced by a grouping transformation that is not independent of the disturbances, or (iii) sampling error introduced by the use of less information in the grouped regression. They also provided new criteria for judging the bias and efficiency of grouping methods. We shall explore their work and Haitovsky's in more detail in Chapter 6.

Hannan (1970a, 1971; 1972) integrated the various approaches to the aggregation problems discussed herein. His extension of Blalock's causal logic is particularly pertinent to future application of this technique to the problems of grouping. The concluding remarks of Hannan's book on aggregation (1971, pp. 116-117) identified the areas where the knowledge of grouping effects was still limited. He called for expanding our understanding of the consequences of estimating individual-level relations from grouped data, the problem of the present inquiry.

CHAPTER 3

ESTIMATION OF THE LINEAR REGRESSION COEFFICIENT FROM GROUPED DATA IN THE SINGLE-REGRESSOR CASE

Chapter 3 focuses on the substantive factors that determine the effects of using grouped data to estimate the relations that exist in data on individuals. For the time being, we consider a linear model with a single regressor X leaving multivariate problems to Chapter 5.

As a point of departure, the methods employed by Prais and Aitchinson (1954) and by Cramer (1964) for examining regression coefficients from grouped data are presented. These methods represent the general econometric approach to the effects of grouping of observations prior to recent work by Haitovsky (1966) and by Feige and Watts (1972). (See Chapter 5 for further discussion of their work.) Potential problems with the earlier econometric approach are cited. The approach of the sociologists Blalock (1964) and Hannan (1970; 1971) is discussed as an alternative to the econometric conceptualization of grouping effects.

The remainder of the chapter is devoted to attempts to develop a mathematical formulation that will account for the grouping effects described by Blalock and Hannan. The concept of a "grouping variable" is introduced to emphasize the relations of the chosen grouping characteristic to the variables of interest. The simple linear model is replaced by a structure which incorporates an interval grouping variable Z. A taxonomy is then generated by considering possible linear relations of Z to the regressor X and of Z to the regressand Y after adjusting for the relation of Z to X. Four categories result when

Z is placed prior to X and Y in the model.

The bias and, where appropriate, the relative efficiency of estimating the regression coefficient (β_{YX}) at the individual level from the gorup means are examined for each taxonomic category. The results indicate that grouping can yield either a biased or an unbiased estimator. The model which incorporates the grouping variable is found to be better suited for treating the problems of data aggregation than the analytical methods of Prais-Aitchinson and Cramer. In particular, the altered model leads to an explicit formulation of the expected bias due to grouping by a variable having specified relations to the X and Y variables.

I. Terminology and Notation

Three types of variables are considered: dependent, independent, and grouping.[1] A dependent variable, or regressand, is an "outcome" or an "effect" in educational investigations. Only the case of a single dependent variable (Y) will be treated.

The independent variables, or regressors, are those the investigator studies as "causes", "determiners", or "predictors" of the variation in the dependent variable. Where there are multiple independent variables, $\underset{\sim}{X}$ [= $(X^{(1)}, \ldots, X^{(k)})$] denotes the k-dimensional vector representation for the complete set, and $X^{(q)}$ refers to any one variable; $q = 1, \ldots, k$. When there is only one independent variable, the superscript will be dropped.

Typically, values of the independent variable are assumed to have

---

[1] Other writers make no formal use of a "grouping variable". Some speak informally of "the method of grouping" [see, e.g., Prais and Aitchinson (1954) and Cramer (1964)].

been established prior to those of the dependent variable. For example, parents' income is logically prior to the educational achievement of their children. Income could be an independent variable since it is not an "outcome" but rather a potential "cause" of student achievement. The model specified for the relations among variables takes this "order" into account and differentiates between causes and outcomes at each step in the chain.

There is a grouping characteristic. In practice, a function of the grouping characteristic, which we shall label as Z or $Z_{(m)}$, assigns the original observations to cardinally numbered groups. The model to be developed in this chapter requires that the values of Z be represented on an interval scale. In Chapter 4 we shall discuss how the model can be used to understand the bias introduced by grouping on a variable that is merely nominal or ordinal.[2]

If a grouping variable is formed from the student characteristic "number of years of mathematics" by use of the rubrics "0-1", "2-3", "4 or more", $Z = 2$ when a student has had more than 1 year of mathematics and less than 4. More formally, when Z is interval, an individual belongs to group i if his value on the grouping characteristic is greater than U_{i-1}, the upper bound in the range for group $i-1$, and less than L_{i+1}, the lower bound of the range for group $i+1$.[3]

---

[2]Ordinal grouping variables can be treated in the same manner as nominal variables. Alternatively, a non-linear transformation can be performed on the categories of the ordinal grouping variable so that it can be treated as interval. In this case each non-linear transformation yields a different grouping variable with different relations to the other study variables.

[3]It is also possible to generate an interval grouping variable from an unordered grouping characteristic by appropriate scaling procedures (e.g., scaling of father's occupation). This option will be discussed in Chapter 4.

Moreover, the value of Z associated with members of group i will be the mean of the group for the grouping characteristic.

Grouping characteristics can (unless binary) be recoded in alternative ways. Thus a single characteristic can yield different "grouping variables" Z. When necessary, $Z_{(m)}$ is used to emphasize that a particular grouping variable forms m groups, perhaps in contrast to an alternative $Z_{(m')}$. A $Z_{(m)}$ may also be contrasted with an alternative $Z'_{(m)}$ which divides the scale on the grouping characteristic differently.

A grouping variable can be generated from an independent variable or a dependent variable, or in some other way. In a study of the relation of parental income (X) to educational achievement (Y), the grouping characteristic could be X and the grouping variable something like "decile rank, in the population, of parents' income". Here, the independent variable and grouping variable are both functions of parental income though their numerical forms differ. X may have been given in the form of actual dollars of income or in terms of income percentiles. The grouping variable Z has the possible values $1, \ldots, 10$; so ten groups can be formed.

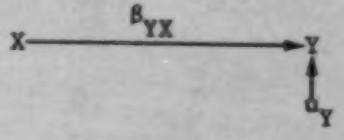
Often, the grouping variable is distinct from both X and Y. For example, observations can be grouped on "father's education", "student's sex", or for that matter, "third letter in student's last name". Indeed, the values of Z can be numbers assigned to persons at random, in which case Z is unrelated to X and Y.

Our models will specify relations among X, Y, and Z, rather than among X, Y, and the grouping characteristic. This is done because observations are actually grouped on a particular Z and two grouping variables generated from the same grouping characteristic can have different relations to X and Y.

A. The Structure Among the Variables

The relations of interest are the structural relations of Y to X . The regression equations represent the presumed underlying structure among the variables. For a given X , there are three possible structural models for the relation between X and Y : (a) X determines Y , (b) Y determines X , (c) there is a reciprocal relation. (There may be other determiners of X and Y denoted by u . When necessary, a subscript is attached to u to identify the variable influenced.) The trivial case of no relation can be ignored.

This investigation concentrates on model (a), which can be represented by the path diagram



The arrows in the diagram indicate the direction of influence; in this case, X determines Y . In a linear model, β_{YX} is the coefficient from the regression of Y on X . u_Y represents all the determiners of Y that are linearly independent of X . Hence u_Y includes errors of measurement in Y , the effects on Y of variables other than X and residuals due to any lack of fit of the linear model. Effects such as u_Y are known as "disturbances" or "disturbance terms". A disturbance u_X could be added prior to X , but this disturbance does not affect the X-Y relation of this model.

The relation depicted in model (a) can be identified by the "structural equation" $Y = \alpha + \beta_{YX}X + u_Y$. This equation specifies that Y can be partitioned into a constant part, a common part due to its linear relation to X , and a residual part, independent of X .

The independent variable X is not partitioned. In factor-analytic terms, two factors can be chosen to account for X and Y, a factor defined by X and a factor for the residual part of Y (that is to say, for Y·X).

B. Notation

We begin with N persons, $p = 1, \ldots, N$. These can be divided among m "groups" on the basis of their Z values. Throughout most of this investigation, the "p" is recoded as ij for clearer designation of group memberships. With the ij notation, Group i contains n_i persons, $n_1 + \ldots + n_i + \ldots + n_m = N$. The labels X_{ij}, Y_{ij}, and Z_{ij} identify the scores of the jth member in the ith group ($i = 1, \ldots, m$; $j = 1, \ldots, n_i$).

Following a standard convention, $\bar{X}_{..}$, $\bar{Y}_{..}$, and $\bar{Z}_{..}$ represent grand means and $\bar{X}_{i.}$, $\bar{Y}_{i.}$, and $\bar{Z}_{i.}$, the means for group i. Under the assumptions made in this investigation, $Z_{ij} = \bar{Z}_{i.}$. The disturbances u_{ij} have group means $\bar{u}_{i.}$. (Later there will be other disturbance terms v and w, to which the same conventions apply.).

Throughout the analyses, population variances and covariances are denoted by σ_X^2, σ_Y^2, σ_{XY}, and so on. Also of interest are population correlation coefficients σ_{XY}, σ_{XZ}, and σ_{YZ} and the coefficient, β_{XZ}, describing the regression of X on Z. The partial regression coefficients $\beta_{YX\cdot Z}$ and $\beta_{YZ\cdot X}$ are important later. In the notation for partials, the effects of the variable placed after the "·" have been controlled when considering the relation of Y to the other regressor.

Additional notation is needed when the sample of persons is only a subset of the population. For the total sample, a sum of squares or sum of cross-products [deviated from the appropriate mean(s)] is identified by $SS_T(\ )$. For example, $SS_T(X)$, denotes the total sum

of squared deviations of X_{ij} from the grand mean:

$$SS_T(X) = \sum_{i=1}^{m} \sum^{n_i}_{j=1} (X_{ij} - \bar{X}_{..})^2$$

Similarly, $SS_B(X)$ represents the between-group sum of squares for X:

$$SS_B(X) = \sum_{i=1}^{m} \sum^{n_i}_{j=1} (\bar{X}_{i.} - \bar{X}_{..})^2$$

$$= \sum_{i=1}^{m} n_i (\bar{X}_{i.} - \bar{X}_{..})^2 \quad .$$

We shall use $SS_W(X)$ to denote a within-group sum of squares:

$$SS_W(X) = \sum_{i=1}^{m} \sum^{n_i}_{j=1} (X_{ij} - \bar{X}_{i.})^2 \quad .$$

The sum of cross-products of X and Y will be denoted by $S(X,Y)$ [$S(\bar{X},\bar{Y})$ for between-group sum of cross-products]. $V(\ )$ and $C(\ ,\ )$ denote the sample variances and covariances -- the sum of squares and sum of cross-products divided by $N-1$, respectively. The sample values of correlation coefficients are represented by r_{XY}, r_{XZ}, and so on.

C. Assumptions About Sampling

In the single-regressor case, our analytical work is based on two sets of assumptions about the sampling of observations. In the simpler case, we take our sample of N persons to be the population of interest. The investigator can then determine β_{YX} from the ungrouped observations.

When observations are grouped on the basis of some Z, the regression analysis performed on the group means $\bar{X}_{i.}$ and $\bar{Y}_{i.}$

(weighted by group size n_i) generates the coefficient $\beta_{\bar{Y}\bar{X}}$ from the population of group means. In this case, where the sample equals the population, our central questions have to do with the adequacy of $\beta_{\bar{Y}\bar{X}}$ as a substitute for β_{YX}; i.e, what is the value of $\beta_{\bar{Y}\bar{X}} - \beta_{YX}$?

Alternatively, we may assume that the persons are a random sample from the population with the constraints that the groups are an exhaustive sample of the values of Z, and the sizes of the groups in the sample are directly proportional to the sizes of the groups in the population. These conditions amount to the implicit assumption that we draw a proportionate stratified sample with strata defined by the values of Z. However, we treat the observations as if they are a simple random sample from the population.

Under the latter sampling assumption (sample ≠ population), the estimator of β_{YX} based on the ungrouped observations is denoted by b_{YX} and its variance over a hypothetical population of independent random samples is denoted by $V(b_{YX})$. The sample estimator of $\beta_{\bar{Y}\bar{X}}$ from the weighted group means is denoted by $B_{\bar{Y}\bar{X}}$, and $V(B_{\bar{Y}\bar{X}})$ represents its variance over samples.

The bias from using grouped data in this case is reflected in the difference between the expected value of $B_{\bar{Y}\bar{X}}$ and β_{YX} [$E(B_{\bar{Y}\bar{X}})-\beta_{YX}$, where expectation is over all i and j]. The difference between $B_{\bar{Y}\bar{X}}$ and b_{YX} provides an estimate of the bias due to grouping.

The relative efficiency of b_{YX} and $B_{\bar{Y}\bar{X}}$ as estimators of β_{YX} is determined by comparing $V(b_{YX})$ to $MSE(B_{\bar{Y}\bar{X}})$, where $MSE(B_{\bar{Y}\bar{X}}) = V(B_{\bar{Y}\bar{X}}) + [E(B_{\bar{Y}\bar{X}}) - \beta_{YX})]^2$. When b_{YX} and $B_{\bar{Y}\bar{X}}$ are unbiased estimators of β_{YX} and $\beta_{\bar{Y}\bar{X}}$, respectively, $MSE(B_{\bar{Y}\bar{X}}) = V(B_{\bar{Y}\bar{X}}) + (\beta_{\bar{Y}\bar{X}} - \beta_{YX})^2$ estimates the mean squared error from estimating β_{YX} from $B_{\bar{Y}\bar{X}}$.

Table 3.1 summarizes the alternative sampling procedures and the measures of precision in estimating from grouped observations under each procedure.

When the data represent a subsample of the population, sampling bias can contribute to the discrepancy between parameter estimates from grouped and ungrouped data. Thus, we potentially confound grouping bias with sampling bias. Treating a proportionate stratified sample as if it were a simple random sample also offers hazards for interpretation. Later, when we talk about bias due to grouping, we do not make the distinction between sampling bias and grouping bias. The combined quantity is attributed to what we call grouping bias or discrepancy.[4]

The assumption of exhaustive sampling of the values of Z causes no special problems when Z is an interval grouping variable based on an interval grouping characteristic. However, whenever the characteristic is nominal, such as school or classroom, the generality of conclusions are restricted by requiring exhaustive sampling. The investigator would like to generalize beyond the classrooms he samples. In any case, the classrooms sampled should at least be randomly representative of some population of interest and lack of representativeness introduces additional bias. This source of bias is also attributed to grouping under the prescribed analytical procedures.

---

[4]Feige and Watts (1972) add specification bias as a third confounding source for the difference between grouped and ungrouped coefficients. In fact, Feige (personal communication) believes that what we call grouping bias is actually specification bias arising from the omission of a relevant variable from the initial model. We do not disagree with this interpretation. However, the generality of the notion of specification bias fails to capture the fact that an investigator may be interested in estimating the simple linear regression coefficient at the individual level, and his problem arises mainly because he must analyze aggregated data.

Table 3.1. Indices of precision of estimates from grouped data as a function of sampling procedure.

Nature of Sample	Description of Sampling Procedure	Measure of Precision: Discrepancy[a]	Measure of Precision: Efficiency
Sample ≡ Population	The sample of N persons constitutes the population.	$\theta = \beta_{\bar{Y}\bar{X}} - \beta_{YX}$	
Sample ≠ Population	N persons are sampled randomly from the population in such a way that all possible values for Z are sampled in proportion to the sizes of the groups in the population.	$d = B_{\bar{Y}\bar{X}} - \beta_{YX}$	$Eff(B,b) = \frac{V(b_{YX})}{\widehat{MSE}(B_{\bar{Y}\bar{X}})}$

[a] E(discrepancy) = Bias

II. Regression Coefficients to be Contrasted

Between-groups regression coefficients can always be estimated from grouped data. For example, assume that in an investigation of the relation between achievement and income, students are grouped on the basis of fathers' education. In this situation, the averages of parental income and student achievement at successive levels of fathers' education and the group sizes n_i become the data for the regression analysis. The investigator can then calculate $B_{\bar{Y}\bar{X}}$, the slope of the regression of group means of achievement on means for income. This is an unbiased estimate of $\beta_{\bar{Y}\bar{X}}$.

However, the purpose of the investigation is to learn about the ungrouped regression coefficient β_{YX}. Our question then is "what is the relation of $B_{\bar{Y}\bar{X}}$ to β_{YX}?". That is, we want to know the conditions under which the slope estimator $(B_{\bar{Y}\bar{X}})$ from the between-groups regression is an unbiased (or possibly just consistent) and efficient estimator of the slope (β_{YX}) from the regression of Y on X using the ungrouped observations. The rest of this inquiry moves toward a statement of these conditions.

III. The Bivariate Case -- Standard Model

We first present a standard statistical model for the relation of Y to X in the ungrouped observations and in the group means. A discussion of the formulation by Cramer (1964) follows[5], with digressions to call attention to important problems of application. Section III.E., in particular, is devoted to the effects of violating

---

[5] We concentrate here on Cramer's bivariate regression analysis rather than the multiple-regression work done by Prais and Aitchinson. The latter will be discussed in more detail in Chapter 5.

assumptions on the Prais-Aitchinson and Cramer conclusions. Finally, we discuss work by Blalock (1964) and Hannan (1970, 1971; 1972) which delineates the effects of grouping in a more realistic manner than the Prais-Aitchinson and Cramer treatments. Throughout this section we deal with the case of subsample from the population.

A. Regression Analysis of the Ungrouped Observations

When a sample of N persons, $p = 1, \ldots, N$, is drawn from the population, the relation between Y_p and X_p is described by the regression equation

$$[3.1] \qquad Y_p = \alpha + \beta_{YX} X_p + u_p \quad .$$

where

$$[3.2] \qquad \beta_{YX} = \frac{\sigma_{YX}}{\sigma_X^2} \quad .$$

One set of assumptions for this model (with random X) is

A1. The X_p are random variables distributed independently of the u_p.

A2. The u_p are independent random disturbances with $E(u_p) = 0$ and $V(u_p) = \sigma_u^2$ for all p.

In this case the least-squares estimator of β_{YX} from the sample of individual data is given by

$$[3.3] \qquad b_{YX} = \frac{C(X_p, Y_p)}{V(X_p)} = \frac{\sum_{p=1}^{N} (X_p - \bar{X}_\cdot)(Y_p - \bar{Y}_\cdot)}{\sum_{p=1}^{N} (X_p - \bar{X}_\cdot)^2} \quad .$$

When [3.2] is substituted for Y_p in [3.3] and the expectation taken, we obtain (by summation over persons)

$$[3.4] \qquad E(b_{YX}) = \beta_{YX} + E\left[\frac{C(X_p, u_p)}{V(X_p)}\right]$$

$$= \beta_{YX} + E\left[\frac{\Sigma(X_p - \bar{X}_\cdot)(u_p - \bar{u}_\cdot)}{\Sigma(X_p - \bar{X}_\cdot)^2}\right] .$$

Since the disturbances u_p and the regressor X_p are assumed to be independent by A1, the second term is zero. So

$$E(b_{YX}) = \beta_{YX} ,$$

and b_{YX} is an unbiased estimator of β_{YX}. (When the u_p are normally distributed, b_{YX} is also the maximum likelihood estimator.)

Under assumptions A1 and A2, the variance of b_{YX} can be shown to be (see, e.g., Goldberger, 1964, p. 267)

$$[3.5] \qquad V(b_{YX}) = E\left[(b_{YX} - \beta_{YX})\right]^2$$

$$= E\left\{\left[\frac{C(X_p, u_p)}{V(X_p)}\right]\right\}^2$$

$$= E\left\{ E\left[\frac{C(X_p, u_p)}{V(X_p)}\right] | X_p \right\}^2$$

$$= \sigma_u^2 E\left[\frac{1}{\sum_{p=1}^{N} (X_p - \bar{X}_\cdot)^2}\right] .$$

If the data satisfy the assumptions on the X_p and u_p and the sampling assumptions, then within the class of linear unbiased estimators of the linear regression coefficient of Y on X, b_{YX} is the estimator with minimum variance (see, e.g., Goldberger, 1964, p. 269).

B. Regression Estimation from Data on Groups

Double subscripts are needed for the sample observations when β_{YX} is estimated from data on groups. The "p" are recoded as "ij" (see Section II.B.). Equation [3.1] becomes

$$[3.1] \qquad Y_{ij} = \alpha + \beta_{YX} X_{ij} + u_{ij} \quad .$$

That is, ij (group i , jth member) replaces p .

We can retain the definitions of b_{YX} and $V(b_{YX})$ given by [3.3] and [3.5], as no change in assumptions has been made about data at the individual level. Note, in particular, that we have assumed sampling of individuals as i,j units, and not sampling of i and of j within i .

In estimating the regression coefficient from group means, any ordering of the groups is ignored. The within-group means, weighted by the number of observations in the group (n_i) , replace the (X_{ij}, Y_{ij}) pairs, and the regression equation relating the $\bar{Y}_{i\cdot}$ to the $\bar{X}_{i\cdot}$ is estimated. We shall hereafter refer to $\beta_{\bar{Y}\bar{X}}$ as the population value of the least-squares coefficient predicting $\bar{Y}_{i\cdot}$ from $\bar{X}_{i\cdot}$, where the means are weighted in proportion to group size in the population. α^* will denote the intercept in this equation.

The relation between $\bar{Y}_{i\cdot}$ and $\bar{X}_{i\cdot}$ is described by the regression equation

$$[3.6] \qquad \bar{Y}_{i\cdot} = \alpha^* + \beta_{\bar{Y}\bar{X}} \bar{X}_{i\cdot} + \bar{u}_{i\cdot} \quad .$$

This equation has the same form as [3.1] where now the group means play the role of "individuals". If the assumption about the u_{ij} holds for the ungrouped observations, the analogous statements also hold for the

grouped observations. (E.g., $E(\bar{u}_{i.}) = 0$.)

We define

$$C(\bar{X}_{i.},\bar{Y}_{i.}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{X}_{i.} - \bar{X}_{..})(\bar{Y}_{i.} - \bar{Y}_{..})}{N-1} ,$$

$$= \frac{\sum_{i=1}^{m} n_i(\bar{X}_{i.} - \bar{X}_{..})(\bar{Y}_{i.} - \bar{Y}_{..})}{N-1} ,$$

and

$$V(\bar{X}_{i.}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{X}_{i.} - \bar{X}_{..})^2}{N-1} ,$$

$$= \frac{\sum_{i=1}^{m} n_i(\bar{X}_{i.} - \bar{X}_{..})^2}{N-1} ,$$

where group means have been weighted by their corresponding n_i . The weighted least-squares estimator, $B_{\bar{Y}\bar{X}}$, of $\beta_{\bar{Y}\bar{X}}$ in [3.6] is then

$$[3.7] \qquad B_{\bar{Y}\bar{X}} = \left[\frac{C(\bar{X}_{i.},\bar{Y}_{i.})}{V(\bar{X}_{i.})}\right] .$$

When [3.6] is substituted for $\bar{Y}_{i.}$ in [3.7], and the expectation taken, we obtain

$$[3.8] \qquad E(B_{\bar{Y}\bar{X}}) = \beta_{\bar{Y}\bar{X}} + E\left[\frac{C(\bar{X}_{i.},\bar{u}_{i.})}{V(\bar{X}_{i.})}\right] .$$

Since $\bar{u}_{i.}$ and $\bar{X}_{i.}$ are independently distributed, the second term is zero, and $B_{\bar{Y}\bar{X}}$ is an unbiased estimator of $\beta_{\bar{Y}\bar{X}}$.

Under the assumptions A1 and A2 , the variance of the grouped estimator is

[3.9] $$V(B_{\bar{Y}\bar{X}}) = E(B_{\bar{Y}\bar{X}} - \beta_{\bar{Y}\bar{X}})^2$$

$$= E\left[\frac{C(\bar{X}_{i.}, \bar{u}_{i.})}{V(\bar{X}_{i.})}\right]$$

$$= \sigma_u^2 E\left[\frac{1}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{X}_{i.} - \bar{X}_{..})^2}\right] .$$

C. Bias and Efficiency of Estimating β_{YX} from Grouped Observations

Though $B_{\bar{Y}\bar{X}}$ is an unbiased estimator of $\beta_{\bar{Y}\bar{X}}$, we are interested in its adequacy as an estimator of β_{YX}, the coefficient from the ungrouped observations. If we let $d = B_{\bar{Y}\bar{X}} - \beta_{YX}$ represent the discrepancy in estimating β_{YX} from $B_{\bar{Y}\bar{X}}$, then the bias from grouping, θ, can be written

[3.10] $$\theta = E(d) = E(B_{\bar{Y}\bar{X}} - \beta_{YX})$$

$$= E(B_{\bar{Y}\bar{X}}) - \beta_{YX}$$

$$= \beta_{\bar{Y}\bar{X}} - \beta_{YX} .$$

Since $E(b_{YX}) = \beta_{YX}$, we may also write

[3.11] $$\theta = E(d) = E(B_{\bar{Y}\bar{X}} - b_{YX}) .$$

According to [3.10], the bias in estimating β_{YX} from $B_{\bar{Y}\bar{X}}$ is zero when the population value of the regression coefficient from grouped data equals the population value of the coefficient from the

ungrouped observations. Furthermore, by [3.11], the bias can be evaluated by comparing the grouped estimator $B_{\bar{Y}\bar{X}}$ with the ungrouped estimator b_{YX}.

We also want to evaluate the efficiency of estimator b_{YX} relative to estimator $B_{\bar{Y}\bar{X}}$ in estimating the regression coefficient from ungrouped data. For the time being, we shall take as our index of the efficiency of the grouped estimator, the ratio of the mean-squared error of b_{YX} to the mean-squared error of $B_{\bar{Y}\bar{X}}$ in estimating β_{YX}; namely,

$$[3.12] \qquad Eff(b_{YX}, B_{\bar{Y}\bar{X}}) = \frac{MSE(b_{YX})}{MSE(B_{\bar{Y}\bar{X}})}$$

$$= \frac{V(b_{YX})}{V(B_{\bar{Y}\bar{X}}) + (\beta_{\bar{Y}\bar{X}} - \beta_{YX})^2}$$

since b_{YX} and $B_{\bar{Y}\bar{X}}$ are unbiased estimators of β_{YX} and $\beta_{\bar{Y}\bar{X}}$, respectively.

When $\beta_{\bar{Y}\bar{X}} = \beta_{YX}$, the efficiency index [3.12] can be written as a ratio of expectations involving the between-group and total sums of squares of X by substitution from [3.5] and [3.9]:

$$[3.13] \qquad Eff(b_{YX}, B_{\bar{Y}\bar{X}}) = \frac{V(b_{YX})}{V(B_{\bar{Y}\bar{X}})}$$

$$= \frac{\sigma_u^2 E[\frac{1}{SS_T(X)}]}{\sigma_u^2 E[\frac{1}{SS_B(X)}]}$$

$$= \frac{E[\frac{1}{SS_T(X)}]}{E[\frac{1}{SS_B(X)}]}$$

From the theorems on the components of variance, the total sum of squares over all N observations can be decomposed in the following

manner:

$$\underset{\text{(Total)}}{SS_T(X)} = \underset{\text{(Between)}}{SS_B(X)} + \underset{\text{(Within)}}{SS_W(X)} \quad ,$$

so that

$$SS_B(X) \leq SS_T(X) \quad .$$

Because all terms are non-negative,

$$\frac{1}{SS_T(X)} \leq \frac{1}{SS_B(X)} \quad ,$$

and

$$E\left[\frac{1}{SS_T(X)}\right] \leq E\left[\frac{1}{SS_B(X)}\right] \quad .$$

Consequently, $Eff(b_{YX}, B_{\overline{YX}}) \leq 1$, and $B_{\overline{YX}}$ is generally less efficient than b_{YX} .

Furthermore, according to [3.13], a grouping procedure that maximizes the between-group sum of squares of the independent variable leads to more efficient estimates. That is, one prefers a procedure which forms groups homogeneous in X_{ij} . So, of those grouping procedures that yield unbiased estimators of β_{YX} , the one which maximizes (minimizes) the between-group (within-group) sum of squares of the independent variable yields the best estimates.

D. Differences from Cramer's Formulation

The analytical work of Cramer (1964) differs in two respects from what has been done so far. First Cramer assumes that the X_{ij} are fixed and given, making the additive disturbance the only random element. Under the assumption of fixed X_{ij} , the sums of squares involving X_{ij} are constants and the expressions for the variances of the estimators can be simplified. That is when the X_{ij} are fixed and given, [3.5]

and [3.9] can be written as

$$V(b_{YX}) = \frac{\sigma_u^2}{SS_T(X)} \quad ,$$

and

$$V(B_{\overline{YX}}) = \frac{\sigma_u^2}{SS_B(X)} \quad ,$$

respectively.

Moreover, when $\beta_{\overline{YX}} = \beta_{YX}$, the efficiency of the grouped estimator becomes

$$Eff(b_{YX}, B_{\overline{YX}}) = \frac{SS_B(X)}{SS_T(X)}$$

$$= \eta_X^2 \quad .$$

This η_X^2 is the correlation ratio.

Here, again, we see that grouped estimators that maximize the between-group variation in the X_{ij}, i.e. that maximize η_X^2, yield the most efficient estimators. Thus, conclusions about the efficiency of estimation are not affected by whether X_{ij} are assumed to be fixed or random.

The other major difference in the Cramer formulation involves his assumptions regarding the sampling of observations and the effects of grouping on the population parameters to be estimated. According to Cramer, the sample of N observations (X_{ij}, Y_{ij}) are "from the outset divided into m groups of n_i observations each, ... The X_{ij} are fixed and given, and the Y_{ij} are repeated samples defined by

[3.1] $$Y_{ij} = \alpha + \beta_{YX} X_{ij} + u_{ij} \quad , \qquad \text{[his equation (1)]}$$

where α and β_{YX} are unknown constants" (Cramer, 1964, p. 235, emphasis added). His assumptions about the disturbances u_{ij} are equivalent to assumptions A1 and A2 above.

Cramer further states that it follows from his equation (1) that

$$\bar{Y}_{i\cdot} = \alpha + \beta_{YX}\bar{X}_{i\cdot} + \bar{u}_{i\cdot} \quad .$$

That is, he assumes that the act of averaging observations within groups does not alter the model assumed to be generating the observations and thus does not affect the parameters that are to be estimated. Thus, from his analysis, we would conclude that $B_{\bar{Y}\bar{X}}$ and b_{YX} as given by our [3.3] and [3.7], respectively, are both unbiased estimates of β_{YX}. (This is also the conclusion reached by Cramer.)

In Section III.B., we state that the equation relating $\bar{Y}_{i\cdot}$ to the $\bar{X}_{i\cdot}$ is

[3.6] $$\bar{Y}_{i\cdot} = \alpha^{*} + \beta_{\bar{Y}\bar{X}}\bar{X}_{i\cdot} + \bar{u}_{i\cdot} \quad ,$$

where parameters α^{*} and $\beta_{\bar{Y}\bar{X}}$ may differ from the parameters α and β_{YX} for the ungrouped observations. This is an important distinction that foreshadows our differing conclusions regarding possible bias from grouping. In the next section we consider how Cramer's assumptions caused him to overlook several plausible grouping procedures that can result in biased estimation.

E. Implications of Assumptions for Equation [3.1]

Not all methods of grouping meet the conditions implied by assumptions A1 and A2 ; neither Cramer nor Prais-Aitchinson notes this explicitly. For example, if the data of students from the school districts of California are used to estimate the regression of student achievement on parental income, it is plausible that the mean distur-

bance will vary according to school district. This would mean that $E(u_{ij}|i,j) = \mu_i$, not necessarily zero or constant. But unless $E(u_{ij} - \bar{u}_{..}) = 0$, we are unable to simplify equations [3.4] and [3.8] when X_{ij} are random variables. That is, if the u_{ij} have a non-zero expectation, b_{YX} and $B_{\bar{Y}\bar{X}}$ are biased estimators of their respective parameters.

Heteroscedasticity and interdependence among the disturbances are other plausible complications. Assumption A1 no longer holds. Under these conditions, the disturbances can be described instead by the equation

$$\text{Cov}\ (u_{ij},u_{i'j'}) = \sigma^2_u \underset{\sim}{\Omega} \quad ,$$

where $\underset{\sim}{\Omega}$ is an $N \times N$ covariance matrix whose off-diagonal cells need not be zero. The elements on the diagonal (variances) may vary according to group (district) and the covariance within a group can be non-zero; that is, $E(u_{ij},u_{ij'}) = \sigma^2_{u_i} \neq 0$.

When heteroscedasticity and interdependence of disturbances are present, least-squares estimators are still unbiased, but they no longer have minimum mean-squared error (cf., Goldberger, 1964, pp. 231-243). this problem can be overcome by transforming the observations so that they satisfy A2 and estimating the parameters from the transformed data. For example, when heteroscedasticity is strictly a function of differences in group size [that is, when $\underset{\sim}{\Omega} = \text{diag}(n_1, \ldots, n_m)$], weighted least-squares procedures using the grouped data perform the necessary adjustments. With more serious complications, as when $\underset{\sim}{\Omega}$ is unknown, econometricians generally place restrictions on $\underset{\sim}{\Omega}$ to permit its estimation from the simple regression model.

The violation of assumption A2 through covariation of regressor

with disturbance has serious consequences for least-squares estimation from grouped data. Covariation between the X_{ij} and the u_{ij} can occur when the regression model is misspecified through the omission of a variable related to both X and Y. It must then operate through the disturbance term. That is, though the regression coefficient β_{YX} from [3.1] is to be estimated, a better specification of the processes at work is

$$Y_{ij} = \alpha + \beta_{YX \cdot W} X_{ij} + \beta_{YW \cdot X} W_{ij} + u^{*}_{ij} \quad ,$$

where W_{ij} is the variable "omitted" from [3.1]. Given the above specification, the least-squares estimator of β_{YX} from the single-regressor model has expectation

$$E(b_{YX}) = \beta_{YX \cdot W} + \beta_{YW \cdot X} b_{WX} \quad ,$$

where b_{WX} is the sample regression coefficient of W on X (cf. Theil, 1957).

The misspecification becomes a problem when β_{YX} is estimated from observations grouped on the omitted variable. By grouping on W (which is at least partially masked by the u_{ij} in [3.1]), the assumption of independence of regressor and disturbance is violated at the grouped level since the W_{ij} are related to both the X_{ij} and the u_{ij}. As a result, $C(\bar{X}_{i.}, \bar{u}_{i.}) \neq 0$ and $B_{\overline{YX}}$ from [3.7] is then a biased estimator of β_{YX}.

Finally, in the present example, the designation of a single constant β_{YX} and the assumptions for the model represent an over-simplification even for the ungrouped observations. Our model does not consider the possibility that the Y-on-X slopes differ because of school district effects. If differential district effects are observed,

the analyst might best examine his data in some multivariate way.

F. Grouping on the Dependent Variable -- Ideas of Blalock and Hannan

Before moving to our own approach to estimation from grouped observations, we point out arguments by Blalock (1964) and Hannan (1971; 1972) that run counter to Cramer and Prais-Aitchinson. Both Blalock and Hannan have argued that systematic grouping methods can yield biased estimators of regression coefficients.

Blalock (1964) based his objection to the "no bias" conclusions of Prais-Aitchinson and Cramer on the findings by Robinson (1950) and others that correlation coefficients are biased by grouping, and on the relation of regression coefficients to the squared correlation coefficients. His reasoning was as follows:

1. Groupings which maximize variation in either X or Y inflate the correlation:

$$r^2_{\bar{X}\bar{Y}} \geq r^2_{XY} \quad .$$

2. According to Prais-Aitchinson and Cramer, grouping on X does not bias the estimate of β_{YX} :

$$E(B_{\bar{Y}\bar{X}}) = E(b_{YX}) = \beta_{YX} \quad .$$

3. The squared correlation r^2_{XY} equals the product of the regression coefficients b_{YX} and b_{XY} .

$$r^2_{XY} = b_{YX}b_{XY} \quad .$$

Similarly,

$$r^2_{\bar{X}\bar{Y}} = B_{\bar{Y}\bar{X}}B_{\bar{X}\bar{Y}} \quad .$$

4. Given the above, it follows that grouping on X inflates the regression coefficient:

$$B_{\bar{X}\bar{Y}} \geq b_{XY} \quad .$$

Blalock's conclusion from the above was that the regression coefficient is inflated when data are grouped on the dependent variable. This apparently contradicts arguments that estimates from grouped observations are always unbiased.

Building on Blalock, Hannan (1972) provided a particularly apt description of how bias can arise through grouping. He argued that bias occurs when observations are grouped on the dependent variable Y. When variation in Y is maximized by ranking observations by their Y values and then grouping "adjacent" observations, observations that have both X values and high u values will be placed in the highest Y groups, assuming β_{YX} is positive. Similarly, observations with both low X values and low u values are placed in the groups lowest on Y. Thus, other determiners of Y are confounded with X so that $C(\bar{X},\bar{u})$ can no longer be expected to equal zero. Hannan stated that this correlation between regressor variable and the disturbance violated the assumptions and was the result of a specification error magnified by grouping. Since the model at the grouped level is misspecified, the least-squares estimators are no longer unbiased.

Blalock's and Hannan's arguments are largely intuitive. In the next section, we present a formal mathematical treatment which supports the contentions of Blalock and Hannan.

IV. A Structural Model for Determining the Effects of Grouping

A systematic procedure is developed for examining the consequences of different methods of grouping observations. The procedure is an extension of the "structural equations" approach by Blalock (1964) and by Hannan (1971; 1972). First an interval grouping variable Z is added to the model of [3.1]. In other words, the rule by which the

individual observations are assigned to groups is treated as a random variable which may be related to other variables in the system. If the grouping variable Z is related to another variable, the structure will specify that Z is prior to that variable. It does not matter that Z may appear to be determined by, say, X in the sense that X would be logically or temporally prior to Z if the three-variable model $Y = f(X,Z)$ were under investigation. We visualize the grouping process as one in which Z can "select" or "force" the observations from the bivariate distribution of X and Y into groups. It is in this sense that Z is prior to X and Y[6].

The equations for the modified structure are presented below for both grouped and ungrouped cases. In addition, general formulas are derived for both grouped and ungrouped coefficients, their estimators, and their variances. Even though we are investigating "a single regressor", we have here a three-variable system where Y can be regressed on X and Z .

Next we consider how the relations of the grouping variable to the other variables affect the usefulness of $B_{\overline{YX}}$ as an estimator of β_{YX} . Problems with regard to the scale and distribution of the variables are set aside for the moment. A taxonomy will be set out such that grouping variables from the modified structure fit into one of several mutually exclusive categories defined by the relations of Z to X and Y .

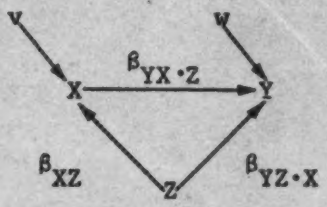
---

[6] This interpretation of Z is in no sense arbitrary. The process of grouping systematically has much in common with the notion of selection. In fact, Lütjohann (personal communication) has suggested that the grouping bias we discuss is essentially selection bias, the result of a manipulated sampling of the observations of X and Y because of their association with Z . Recent work by Goldberger (1972) on selection bias in evaluating treatment effects with non-random sampling also hints at the connection.

As later sections will demonstrate, the use of this taxonomic structure enables the investigator to reject many potential grouping variables by examining their matrix of correlations and partial correlations with the main variables in the study.

A. Structure with Z prior to X and Y

The path diagram for the structure when Z is prior to X and Y is



In this diagram, v is the disturbance term representing all determiners of X that are not linearly related to Z , and w is the disturbance term representing all determiners of Y that are not linearly related to X or Z . $\beta_{YX \cdot Z}$, $\beta_{YZ \cdot X}$, and β_{XZ} are the path regression coefficients.

The equations corresponding to the structure with Z incorporated can be written

[3.14a) $$Y = \alpha + \beta_{YX \cdot Z} X + \beta_{YZ \cdot X} Z + w \quad ,$$

[3.14b) $$X = \lambda + \beta_{XZ} Z + v \quad .$$

We recall that $\beta_{YX \cdot Z}$, $\beta_{YZ \cdot X}$, and β_{XZ} refer to regression parameters in a system with several variables. Even though we include the grouping variable Z , this is an equation at the individual level; every person has a Z_p . w and v are disturbance terms with zero expected values for all persons. w is assumed to be independent of X, Z , and v ; and v is assumed to be independent of Z . We further assume that

both disturbance terms are homoscedastic (i.e., for any two persons, $\sigma^2_{w_1} = \sigma^2_{w_2} = \sigma^2_w$, $\sigma^2_{v_1} = \sigma^2_{v_2} = \sigma^2_v$) and independent. (This implies that for any two persons, $\sigma_{w_1 w_2} = 0$) and $\sigma_{v_1 v_2} = 0$.

Although we again write α for intercept term in [3.14a], its value may differ from that in earlier equations. We let λ represent the intercept term in the second equation of the structural system.

Equation [3.14b] can be substituted into [3.14a] to obtain a single equation for the regression of Y on Z and v :

$$[3.15] \quad Y = (\alpha + \beta_{YX \cdot Z}\lambda) + (\beta_{YX \cdot Z}\beta_{XZ} + \beta_{YZ \cdot X})Z + \beta_{YX \cdot Z}v + w \quad .$$

Equation [3.15] is actually a reparameterization of [3.1] where X has been divided into two parts -- the part predictable from the grouping variable Z and a residual part v . Equations like [3.15] are generally called "reduced-form" equations. This means that [3.15] is in a form that cannot be reduced further by substitution of other equations from the structural system. Later on, we use reduced-form expressions to simplify our analytical work.

In Table 3.2, expressions for the population variances and covariances of the variables in equations [3.14a] and [3.14b] are provided. The corresponding reduced-form versions are enclosed in brackets.

The regression coefficient relating Y to X -- the ratio of σ_{XY} to σ^2_X as given in Table 3.2 -- is equivalent to the coefficient given by [3.1]. As can be seen, that ratio involves the three regression coefficients ($\beta_{YX \cdot Z}$, $\beta_{YZ \cdot X}$, and β_{XZ}) and the variances σ^2_Z , σ^2_v , and σ^2_w .

Table 3.2. Covariance matrix for variables in equations [3.14a] and [3.14b] (Reduced forms in brackets).

Variable	Y	X	Z	x	v
Y	$\beta^2_{YX\cdot Z}\sigma^2_X + \beta^2_{YZ\cdot X}\sigma^2_Z + \beta_{YX\cdot Z}\beta_{YZ\cdot X}\sigma_{XZ} + \sigma^2_w \equiv$ $[(\beta_{YX\cdot Z}\beta_{XZ} + \beta_{YZ\cdot X})^2\sigma^2_Z + \beta^2_{YX\cdot Z}\sigma^2_v + \sigma^2_w]$				
X	$\beta_{YX\cdot Z}\sigma^2_X + \beta_{YZ\cdot X}\sigma_{XZ} \equiv$ $[\beta_{XZ}(\beta_{YZ\cdot X} + \beta_{YX\cdot Z}\beta_{XZ})\sigma^2_Z + \beta^2_{YX\cdot Z}\sigma^2_v]$	$\sigma^2_X \equiv$ $[\beta^2_{XZ}\sigma^2_Z + \sigma^2_v]$			
Z	$\beta^2_{YZ\cdot X}\sigma^2_Z + \beta_{YX\cdot Z}\sigma_{XZ} \equiv$ $[(\beta_{YX\cdot Z}\beta_{XZ} + \beta_{YZ\cdot X})^2\sigma^2_Z]$	$\sigma_{XZ} \equiv$ $[\beta_{XZ}\sigma^2_Z]$	σ^2_Z		
w	σ^2_w	0	0	σ^2_w	
v	$[\beta_{YX\cdot Z}\sigma^2_v]$	σ^2_v	0	0	σ^2_v

By substitution (of the reduced-form expressions from Table 3.2), we get

$$[3.16] \qquad \beta_{YX} = \frac{\sigma_{YX}}{\sigma_X^2}$$

$$= \frac{\beta_{XZ}(\beta_{YZ\cdot X} + \beta_{YX\cdot Z}\beta_{XZ})\sigma_Z^2 + \beta_{YX\cdot Z}\sigma_v^2}{\beta_{XZ}^2\sigma_Z^2 + \sigma_v^2}$$

$$= \beta_{YX\cdot Z} + \beta_{YZ\cdot X}\beta_{XZ}\left(\frac{\sigma_Z^2}{\sigma_X^2}\right) .$$

B. Revised Structure for Weighted Group Means

The structural equations for the means of groups with uniform Z can be written as

$$[3.17a] \qquad \bar{Y} = \alpha + \beta_{YX\cdot Z}\bar{X} + \beta_{YZ\cdot X}\bar{Z} + \bar{w} ,$$

$$[3.17b] \qquad \bar{X} = \lambda + \beta_{XZ}\bar{Z} + \bar{v} .$$

These equations are the same as [3.14a] and [3.14b] except that grouped quantities have been substituted for their ungrouped counterparts. In addition to the intercepts, there are still six parameters, $\beta_{YX\cdot Z}$, $\beta_{YZ\cdot X}$, β_{XZ}, $\sigma_{\bar{Z}}^2$, $\sigma_{\bar{v}}^2$, and $\sigma_{\bar{w}}^2$. Note that we specify the same regression parameters as in [3.14], since averaging observations within groups does not alter the model underlying the generation of observations. (This is analogous to Cramer's assumption discussed in Section III.D though now we operate with a more correctly specified model.)

Table 3.3 contains the population values for the variances and covariances of the variables in equations [3.17a] and [3.17b]. The reduced forms are again enclosed in brackets.

Table 3.3. Covariance matrix for variables in equations [3.17a] and [3.17b] (Reduced forms in brackets)

Variable	$\bar{Y}$	$\bar{X}$	$\bar{Z}$	$\bar{x}$	$\bar{v}$
$\bar{Y}$	$\beta^2_{YX\cdot Z}\sigma^2_{\bar{X}} + \beta^2_{YZ\cdot X}\sigma^2_{\bar{Z}} + \beta_{YX\cdot Z}\beta_{YZ\cdot X}\sigma_{\bar{X}\bar{Z}} + \sigma^2_{\bar{w}} \equiv$ $[(\beta_{YX\cdot Z}\beta_{XZ} + \beta_{YZ\cdot X})^2\sigma^2_{\bar{Z}} + \beta^2_{YX\cdot Z}\sigma^2_{\bar{v}} + \sigma^2_{\bar{w}}]$				
$\bar{X}$	$\beta_{YX\cdot Z}\sigma^2_{\bar{X}} + \beta_{YZ\cdot X}\sigma_{\bar{X}\bar{Z}} \equiv$ $[\beta_{XZ}(\beta_{YZ\cdot X} + \beta_{YX\cdot Z}\beta_{XZ})\sigma^2_{\bar{Z}} + \beta^2_{YX\cdot Z}\sigma^2_{\bar{v}}]$	$\sigma^2_{\bar{X}} \equiv$ $[\beta^2_{XZ}\sigma^2_{\bar{Z}} + \sigma^2_{\bar{v}}]$			
$\bar{Z}$	$\beta^2_{YZ\cdot X}\sigma^2_{\bar{Z}} + \beta_{YX\cdot Z}\sigma_{\bar{X}\bar{Z}} \equiv$ $[(\beta_{YX\cdot Z}\beta_{XZ} + \beta_{YZ\cdot X})^2\sigma^2_{\bar{Z}}]$	$\sigma_{\bar{X}\bar{Z}} \equiv$ $[\beta_{XZ}\sigma^2_{\bar{Z}}]$	$\sigma^2_{\bar{Z}}$		
$\bar{w}$	$\sigma^2_{\bar{w}}$	0	0	$\sigma^2_{\bar{w}}$	
$\bar{v}$	$[\beta_{YX\cdot Z}\sigma^2_{\bar{v}}]$	$\sigma^2_{\bar{v}}$	0	0	$\sigma^2_{\bar{v}}$

58

By substitution of the reduced-form expressions from Table 3.3, the regression coefficient relating $\bar{Y}$ to $\bar{X}$ -- the ratio of $\sigma^2_{\bar{X}\bar{Y}}$ to $\sigma^2_{\bar{X}}$ -- can be written as

$$[3.18] \qquad \beta_{\bar{Y}\bar{X}} = \frac{\sigma_{\bar{Y}\bar{X}}}{\sigma^2_{\bar{X}}}$$

$$= \beta_{YX \cdot Z} + \beta_{YZ \cdot X}\beta_{XZ}\left(\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{\bar{X}}}\right) .$$

Comparing [3.16] and [3.18], we see that $\beta_{\bar{Y}\bar{X}}$ and β_{YX} differ in that between-group variances replace total variances. When our sample constitutes the entire population (the first case in Table 3.1), the discrepancy, or bias, θ, can be found by substituting from [3.16] and [3.18] for the appropriate terms in [3.10]:

$$[3.19] \qquad \theta = \beta_{\bar{Y}\bar{X}} - \beta_{YX}$$

$$= \beta_{YZ \cdot X}\beta_{XZ}\left(\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{\bar{X}}} - \frac{\sigma^2_{Z}}{\sigma^2_{X}}\right) .$$

C. Estimator of β_{YX} from Individual Data

Under the modified structure, a simple random sample of $N(\sum_{i=1}^{m} n_i)$ observations is drawn from the trivariate distribution $f(X_{ij}, Y_{ij}, Z_{ij})$ generated by [3.14a] and [3.14b]. The sample regression estimator of β_{YX} is given by

$$[3.10] \qquad b_{YX} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..})}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2}$$

$$= \frac{\Sigma xy}{\Sigma x^2} ,$$

where $x = X_{ij} - \bar{X}_{..}$ and $y = Y_{ij} - \bar{Y}_{..}$ are deviation scores and summation is over all N persons.

Equation [3.20] is a double-scripted version of equation [3.3]. An expression for the expected value of b_{YX} in terms of parameters of the modified structure is found by substituting [3.14a] for Y_{ij} in [3.20] (all variables in deviation form) and taking the expectation:

$$[3.21] \quad E(b_{YX}) = E\left[\frac{\Sigma xy}{\Sigma x^2}\right]$$

$$= E\left[\frac{\Sigma x(\beta_{YX\cdot Z}x + \beta_{YZ\cdot X}z + w)}{\Sigma x^2}\right]$$

$$= \beta_{YX\cdot Z} + \beta_{YZ\cdot X}E\left(\frac{\Sigma xz}{\Sigma x^2}\right) + E\left(\frac{\Sigma xw}{\Sigma x^2}\right)$$

$$= \beta_{YX\cdot Z} + \beta_{YZ\cdot X}E\left(\frac{\Sigma xz}{\Sigma x^2}\right) \quad .$$

Equation [3.21] is in a form that cannot be simplified without additional assumptions since, by [3.14b], x and z may be related. We can, however, examine the asymptotic properties of the expression under the conditions that both Σxz and Σx^2 exist and Σx^2 is non-zero. By the Strong Law of Large Numbers [7],

$$\text{plim}\left[\frac{\Sigma xz}{\Sigma x^2}\right] = \text{plim}\left(\frac{N}{\Sigma x^2}\right)\text{plim}\left(\frac{\Sigma xz}{N}\right) \quad ,$$

where plim denotes the probability limit ($\lim_{N \to \infty}$) of the enclosed

---

[7] I am indebted to Professor Julius Blum for pointing out that the Strong Law of Large Numbers is useful in this situation.

expressions. The right-hand side can be further simplified since

$$\text{plim}\left(\frac{N}{\Sigma x^2}\right) = \frac{1}{\sigma_X^2} \quad ,$$

and

$$\text{plim}\left(\frac{\Sigma xz}{N}\right) = \text{plim}\ \frac{\Sigma(\beta_{XZ}z + vz)}{N}$$

$$= \text{plim}\ \beta_{XZ}\frac{\Sigma z^2}{N} + \text{plim}\ \frac{\Sigma vz}{N} = \beta_{XZ}\sigma_Z^2 + \sigma_{vz}$$

$$= \beta_{XZ}\sigma_Z^2$$

since v and z are independent.

Therefore,

[3.22] $$\text{plim}(b_{YX}) = \beta_{YX\cdot Z} + \beta_{YZ\cdot X}\beta_{XZ}\left(\frac{\sigma_Z^2}{\sigma_X^2}\right) \quad ,$$

where, as expected, the right-hand side of [3.22] is the same as the right-hand side of [3.16].

The variance of b_{YX} under the modified structure can be written as

[3.23] $$V(b_{YX}) = E[b_{YX} - E(b_{YX})]^2$$

$$= E\left\{\frac{\Sigma xy}{\Sigma x^2} - \left[\beta_{YX\cdot Z} + \beta_{YZ\cdot X}E\left(\frac{\Sigma xz}{\Sigma x^2}\right)\right]\right\}^2 \quad .$$

Substituting [3.10] and [3.21] in [3.23].

$$V(b_{YX}) = E\left\{\left[\beta_{YX\cdot Z} + \beta_{YZ\cdot X}\left(\frac{\Sigma xz}{\Sigma x^2}\right) + \left(\frac{\Sigma xw}{\Sigma x^2}\right)\right]\right.$$

$$\left. - \left[\beta_{YX\cdot Z} + \beta_{YZ\cdot X}E\left(\frac{\Sigma xz}{\Sigma x^2}\right)\right]\right\}^2 \quad ,$$

and, after substituting the deviation form of [3.14a] for y ,

$$V(b_{YX}) = \Sigma \left\{ \beta_{YZ\cdot X} \left[\frac{\Sigma xz}{\Sigma x^2} - E\left(\frac{\Sigma xz}{\Sigma x^2} \right) \right] + \frac{\Sigma xw}{\Sigma x^2} \right\}^2 .$$

By expanding the right-hand side and applying the assumptions that w is independent of x and z and $E(w) = 0$, [3.23] can be further reduced to

$$V(b_{XY}) = \beta_{YZ\cdot X} E\left[\frac{\Sigma xz}{\Sigma x^2} - E\left(\frac{\Sigma xz}{\Sigma x^2} \right) \right]^2 + \left(\frac{\Sigma xw}{\Sigma x^2} \right)^2 .$$

The last term in the above expression is equal to $\sigma_w^2 E[\frac{1}{SS_T(X)}]$ by the same reasoning we used to derive $V(b_{YX})$ (Equation [3.5]) in Section III.A. Also for the time being, we shall use the fact that $\frac{\Sigma xz}{\Sigma x^2}$ is the expression for the least-squares estimator b_{ZX} to simplify the equation for the variance:

$$V(b_{YX}) = \beta_{YZ\cdot X}^2 V(b_{ZX}) + \sigma_w^2 E\left[\frac{1}{SS_T(X)} \right] .$$

It is difficult to simplify [3.23] further because x is a function of both z and v under the most general conditions. Later, we examine the $V(b_{YX})$ under conditions where Z is assumed to be unrelated to X, to $Y\cdot X$, or to both. In these cases, the expression for the variance of the estimator of β_{YX} from ungrouped data can be simplified.

D. Estimator from Grouped Data

The Y_{ij} and X_{ij}, from the sample of N observations drawn from the trivariate distribution $f(X_{ij}, Y_{ij}, Z_{ij})$, are grouped on the basis of the values of Z_{ij}. Each observation is then replaced by the group mean corresponding to its Z_{ij} value; that is, $\bar{X}_{i\cdot}$ replaces X_{ij} and $\bar{Y}_{i\cdot}$ replaces Y_{ij}. In this treatment, $Z_{ij} = \bar{Z}_{i\cdot}$ so that $\sigma_{\bar{Z}}^2 = \sigma_Z^2$. Furthermore, we assume that the group sizes in the sample -- the n_i's --

are proportional to the group sizes in the population so that bias has not been introduced through non-proportionate sampling from groups.

The equation for the sample regression coefficient $B_{\bar{Y}\bar{X}}$ can be written as:

$$[3.24] \qquad B_{\bar{Y}\bar{X}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2} = \frac{\Sigma\bar{x}\bar{y}}{\Sigma\bar{x}^2}$$

where lower-case letters denote deviations of group means from the grand means of the sample and summation is over all N observations. Though written in a different manner, equation [3.24] is simply the double-scripted version of [3.7].

We now follow the same procedures used in Section IV.C for ungrouped data in order to find the expected value of the sample estimator from grouped data under the modified structure. Substituting the deviation form of [3.14a] for $\bar{y}$ in [3.24] and taking the expectation, we obtain

$$[3.25] \qquad E(B_{\bar{Y}\bar{X}}) = E\left(\frac{\Sigma\bar{x}\bar{y}}{\Sigma\bar{x}^2}\right)$$

$$= E\left[\frac{\Sigma\bar{x}(\beta_{YX\cdot Z}\bar{x} + \beta_{YZ\cdot X}\bar{z} + \bar{w})}{\Sigma\bar{x}^2}\right]$$

$$= \beta_{YX\cdot Z} + \beta_{YZ\cdot X}E\left(\frac{\Sigma\bar{x}\bar{z}}{\Sigma\bar{x}^2}\right) + E\left(\frac{\Sigma\bar{x}\bar{w}}{\Sigma\bar{x}^2}\right)$$

$$= \beta_{YX\cdot Z} + \beta_{YZ\cdot X}E\left(\frac{\Sigma\bar{x}\bar{z}}{\Sigma\bar{x}^2}\right)$$

since x and w (and $\bar{x}$ and $\bar{w}$) are assumed to be independent and $E(\bar{w}) = E(w) = 0$.

By the same reasoning used to derive [3.22], it can be shown that asymptotically

$$[3.26] \qquad \operatorname{plim}(B_{\bar{Y}\bar{X}}) = \beta_{YX\cdot Z} + \beta_{YZ\cdot X}\beta_{XZ}\left(\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{\bar{X}}}\right) .$$

The right-hand side of [3.26] is the same as the right-hand side of [3.18].

An expression for the variance of $B_{\bar{Y}\bar{X}}$ under the modified structure is found in the same fashion as $V(b_{YX})$ in [3.23]. It can be shown that

$$[3.27] \quad V(B_{\bar{Y}\bar{X}}) = E[B_{\bar{Y}\bar{X}} - E(B_{\bar{Y}\bar{X}})]^2$$

$$= \beta_{YZ\cdot X}^2 E\left[\frac{\Sigma\bar{x}\bar{z}}{\Sigma\bar{x}^2} - E\ \frac{\Sigma\bar{x}\bar{z}}{\Sigma\bar{x}^2}\right]^2 + \sigma_w^2 E\left[\frac{1}{SS_B(X)}\right]$$

$$= \beta_{YZ\cdot X}^2 V(B_{\bar{Z}\bar{X}}) + \sigma_w^2 E\left[\frac{1}{SS_B(X)}\right] \quad ,$$

where $B_{\bar{Z}\bar{X}}$ is the least-squares estimator from the regression of $\bar{Z}$ on $\bar{X}$ over all N persons.

The only differences between the equations for grouped and ungrouped coefficients ([3.16] and [3.18]), their sample estimators ([3.21] and [3.25]; also [3.22] and [3.26] for the asymptotic expressions), and sample variances ([3.23] and [3.27]) are that sums of squares and variances of the group means of Z and X replace the sums of squares and variances of the corresponding ungrouped observations. And, since $\sigma_{\bar{Z}}^2 = \sigma_Z^2$ and $SS_B(Z) = SS_T(Z)$ under the modified structure, the only substantive changes involve variation of the independent variable.

b_{YX} and $B_{\bar{Y}\bar{X}}$ have been shown to be asymptotically unbiased estimators of β_{YX} and $\beta_{\bar{Y}\bar{X}}$ respectively, but the investigator wants to estimate β_{YX} from $\beta_{\bar{Y}\bar{X}}$ (when the sample equals the population) or from $B_{\bar{Y}\bar{X}}$. In Section V.B we shall identify the conditions under which $\beta_{\bar{Y}\bar{X}} = \beta_{YX}$ and $B_{\bar{Y}\bar{X}}$ is an unbiased estimator of β_{YX}.

E. A Taxonomy for Classifying Grouping Variables

A taxonomy for comparing grouping variables can be formed by setting various combinations of $\beta_{YZ\cdot X}$ and β_{XZ} in [3.14a] and [3.14b] equal to zero. The categories of the taxonomy reflect different sets of constraints on the relations of Z to Y and X. Four categories of

grouping variables can be distinguished:

I. Z is directly related to both X and $Y \cdot X$ ($\beta_{YZ \cdot X} \neq 0$, $\beta_{XZ} \neq 0$) .

II. Z is directly related to $Y \cdot X$ but not to X ($\beta_{YZ \cdot X} \neq 0$, $\beta_{XZ} = 0$) .

III. Z is directly related to X but not to $Y \cdot X$ ($\beta_{YZ \cdot X} = 0$, $\beta_{XZ} \neq 0$) .

IV. Z is not related to either X or $Y \cdot X$ ($\beta_{YZ \cdot X} = 0$, $\beta_{XZ} = 0$) .

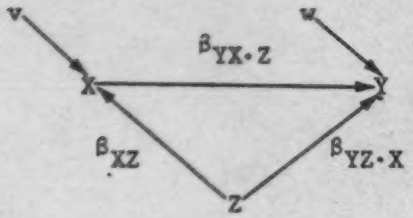
Figure 3.1 presents the path diagrams corresponding to the categories of the taxonomy.

The categories of the taxonomy include all possible linear relations linking prior grouping variables to the regression of Y on X . Certain of these categories represent broader classes of variables. For instance, any random grouping procedure will satisfy the conditions for Category IV. Groupir › on the regressor X is a special case of Category III. Most systematic grouping variables belong to Category I. Grouping on the dependent variable Y is a special case of Category I. Any grouping variable can be uniquely categorized if the variances and covariances of X , Y , and Z are known.

Under certain conditions discussed in Chapter 1, however, no ungrouped estimate of σ_{YX} is available. To see this, suppose that data on X and Z are collected anonymously on occasion 1 and data on Y and Z are collected on occasion 2. Then σ_X , σ_Y , σ_X , σ_{XZ} , and σ_{YZ} can be estimated directly from the data. But there is no natural way to pair X and Y scores, and σ_{YX} and thus β_{YX} cannot be estimated directly. When this occurs, the investigator can estimate β_{YZ} and β_{XZ} , but not $\beta_{YZ \cdot X}$. He can often guess whether $\beta_{YZ \cdot X}$ is

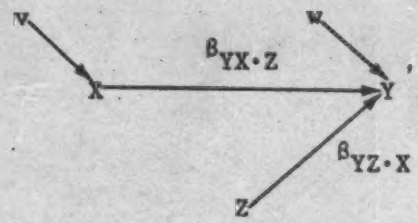
(a) Category I

$\beta_{XZ \cdot X} \neq 0$, $\beta_{XZ} \neq 0$



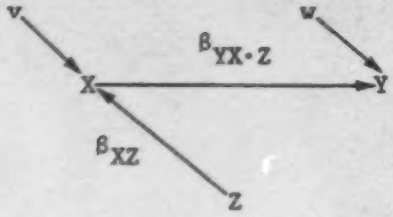
(b) Category II

$\beta_{YZ \cdot X} \neq 0$, $\beta_{XZ} = 0$



(c) Category III

$\beta_{YZ \cdot X} = 0$, $\beta_{XZ} \neq 0$



(d) Category IV

$\beta_{YZ \cdot X} = 0$, $\beta_{XZ} = 0$

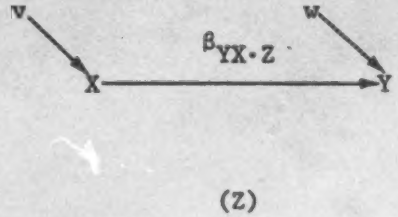


Figure 3.1. Path Diagrams Corresponding to the Categories of the Taxonomy.

non-zero, and by doing so can judge whether grouping by Z will yield unbiased and efficient estimates of $\beta_{YZ\cdot X}$. In Chapter 6, we shall offer suggestions for grouping when σ_{YX} is unknown (cf., Section II.C of Chapter 6).

V. Bias and Efficiency as a Function of Taxonomic Category

We examine how the relations specified for the taxonomic categories can affect the bias and efficiency of the regression estimates from grouped data. First, the general formulas for bias and efficiency from Section III.C are developed for the modified structure. Then the implications of this formulation are considered for each category separately.

A. Bias and Efficiency Formulas

In Section IV.B, we presented the following expression for the discrepancy (bias) that results from grouping when the sample constitutes the population:

[3.19] $$\theta = \beta_{\bar{Y}\bar{X}} - \beta_{YX}$$

$$= \beta_{YZ\cdot X}\beta_{XZ}\left(\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{\bar{X}}} - \frac{\sigma^2_Z}{\sigma^2_X}\right) .$$

When the data are a subsample from the population, the asympotic expression for the bias from grouping found by comparing $\text{plim}(B_{\bar{Y}\bar{X}})$ (Equation [3.26]) with $\text{plim}(b_{YX})$ (Equation [3.22]) has the same form as [3.19]:

[3.28] $$\text{plim}(d) = \text{plim}(B_{\bar{Y}\bar{X}}) - \text{plim}(b_{YX})$$

$$= \beta_{YZ\cdot X}\beta_{XZ}\left(\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{\bar{X}}} - \frac{\sigma^2_Z}{\sigma^2_X}\right) .$$

Also, comparing [3.25] with [3.21], the expectation of the difference between $B_{\bar{Y}\bar{Z}}$ and b_{YX} is given by

$$[3.29] \quad E(d) = E(B_{\bar{Y}\bar{X}}) - \beta_{YX}$$

$$= E(B_{\bar{Y}\bar{X}}) - E(b_{YX})$$

$$= \left[\beta_{YX\cdot Z} + \beta_{YZ\cdot X} E\left(\frac{\Sigma\bar{x}\bar{z}}{\Sigma\bar{x}^2}\right)\right] - \left[\beta_{YX\cdot Z} + \beta_{YZ\cdot X} E\left(\frac{\Sigma xz}{\Sigma x^2}\right)\right]$$

$$= \beta_{YZ\cdot X} E\left(\frac{\Sigma\bar{x}\bar{z}}{\Sigma\bar{x}^2} - \frac{\Sigma xz}{\Sigma x^2}\right) .$$

Since $z_{ij} = \bar{z}_{i\cdot}$,

$$\Sigma xz = \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} z_{ij}$$

$$= \sum_{i=1}^{m} \bar{z}_{i\cdot} \sum_{j=1}^{n_i} x_{ij}$$

$$= \sum_{i=1}^{m} n_i \bar{x}_{i\cdot} \bar{z}_{i\cdot}$$

$$= \Sigma\bar{x}\bar{z} .$$

Similarly,

$$\Sigma vz = \Sigma\bar{v}\bar{z} .$$

So [3.29] can be written as

$$[3.29] \quad E(d) = \beta_{YZ\cdot X} E\left(\frac{\Sigma\bar{x}\bar{z}}{\Sigma\bar{x}^2} - \frac{\Sigma xz}{\Sigma x^2}\right)$$

$$= \beta_{YZ\cdot X} E\left[(\Sigma xz)\left(\frac{1}{\Sigma\bar{x}^2} - \frac{1}{\Sigma x^2}\right)\right]$$

$$= \beta_{YZ\cdot X} E\left[(\beta_{XZ}\Sigma z^2 + \Sigma zv)\left(\frac{1}{\Sigma\bar{x}^2} - \frac{1}{\Sigma x^2}\right)\right]$$

$$= \beta_{YZ\cdot X}\beta_{XZ}E\left[(\Sigma z^2)\left(\frac{1}{\Sigma\bar{x}^2} - \frac{1}{\Sigma x^2}\right)\right] + \beta_{YZ\cdot X}E\left[(\Sigma zv)\left(\frac{1}{\Sigma\bar{x}^2} - \frac{1}{\Sigma x^2}\right)\right]$$

$$= \beta_{YZ\cdot X}\beta_{XZ}E\left[(\Sigma z^2)\left(\frac{\Sigma x^2 - \Sigma\bar{x}^2}{\Sigma\bar{x}^2\Sigma x^2}\right)\right] + \beta_{YZ\cdot X}E\left[(\Sigma xz)\left(\frac{\Sigma x^2 - \Sigma\bar{x}^2}{\Sigma\bar{x}^2\Sigma x^2}\right)\right].$$

With the exception of the last term, [3.29] now has the same form and components as [3.19] and [3.28]. In each case, the bias term has essentially the same straightforward interpretation if the between-group and total variation of X are both non-zero. The grouping of observations leads to biased estimation if all three of the following conditions hold:

(a) The grouping variable Z has a direct relation to X ($\beta_{XZ} \neq 0$) .

(b) The grouping variable Z has a direct relation to $Y\cdot X$ ($\beta_{YZ\cdot X} \neq 0$) .

(c) The ratio of the between-group variation of Z to the between-group variation of X does not equal the ratio of the total variation of Z to the total variation of X .

Furthermore, since Z has been defined so that $Z_{ij} = \bar{Z}_{i\cdot}$, we can rewrite [3.19] and [3.28] as

$$[3.19'] \quad \theta = E(d) = \beta_{YZ\cdot X}\beta_{XZ}\sigma_Z^2\left(\frac{\sigma_{\bar{X}}^2 - \sigma_X^2}{\sigma_{\bar{X}}^2\sigma_X^2}\right) \qquad \text{(sample} \equiv \text{population)}$$

and

$$[3.28'] \quad \text{plim}(d) = \beta_{YZ\cdot X}\beta_{XZ}\sigma_Z^2\left(\frac{\sigma_X^2 - \sigma_{\bar{X}}^2}{\sigma_{\bar{X}}^2\sigma_X^2}\right) \qquad \text{(sample} \neq \text{population)}$$

Thus, condition (c) can be restated as

(c') The between-group variation of X does not equal the total variation of X .

Other things being equal, the magnitude of the bias from grouping increases directly as the relation of Z to X or Y·X increases or as the variation of X is reduced by grouping. These three conditions are not independent; in the next section, we explore some ramifications of their interrelation.

The formula for the efficiency of b_{YX} relative to $B_{\bar{\bar{Y}}\bar{\bar{X}}}$ as an estimator of β_{YX} can be found by substituting from [3.19'], [3.23], and [3.27] into [3.12]:

$$[3.30] \quad Eff(b_{YX}, B_{\bar{\bar{Y}}\bar{\bar{X}}}) = \frac{MSE(b_{YX})}{MSE(B_{\bar{\bar{Y}}\bar{\bar{X}}})}$$

$$= \frac{V(b_{YX})}{V(B_{\bar{\bar{Y}}\bar{\bar{X}}}) + (\beta_{\bar{\bar{Y}}\bar{\bar{X}}} - \beta_{YX})^2}$$

$$= \frac{\beta_{YZ\cdot X} V(b_{XZ}) + \sigma_w^2 E\left[\frac{1}{SS_T(X)}\right]}{\left\{\left[\beta^2_{YZ\cdot X} V(B_{\bar{Z}\bar{X}}) + \sigma_w^2 E\left(\frac{1}{SS_B(X)}\right)\right] + \left[\beta^2_{YZ\cdot X}\beta^2_{XZ}\sigma_Z^2\left(\frac{\sigma_X^2 - \sigma_{\bar{X}}^2}{\sigma_{\bar{X}}^2\sigma_X^2}\right)\right]\right\}}$$

For certain categories, this complicated expression will simplify greatly as Section V.C will show.

B. Examination of Bias for Each Category

Equations [3.19'], [3.28'] and [3.29] can now be used to examine each category of grouping variables for bias. The taxonomic categories are considered in order.

1. Category I -- Z directly related to both X and Y·X .
($\beta_{YZ\cdot X} \neq 0$, $\beta_{XZ} \neq 0$).

Category I includes all grouping variables which have direct relations to both X and Y·X . An obvious example is that scholastic aptitude (Z) may be related to achievement (Y) and to student academic interests (X) .

A more complicated example occurs when two distinct classifications are made on the same achievement measure; for example, define Y as the observed score on achievement and Z as the decile rank on achievement. Thus Z will most likely be a Category I variable. The broader classification for Z creates a measure whose correlation with Y is other than 1.0 or 0 after X is partialled out. If Y is linearly related to X , Z will also be related to X .

In general, the slope estimated from data grouped on Category I variable is a biased estimate of β_{YX} . The magnitude of this bias is given exactly by [3.19'] for known values of $\beta_{YZ\cdot X}$, β_{XZ} , σ_X^2 , and $\sigma_{\bar{X}}^2$ and can be approximated by [3.28'] and [3.29] when the sample does not equal the population.

Thus, when Z is a Category I variable, bias is given by the general equations:

[3.19'] $$\theta = \beta_{YZ\cdot X}\beta_{XZ}\sigma_Z^2\left(\frac{\sigma_X^2 - \sigma_{\bar{X}}^2}{\sigma_{\bar{X}}^2\sigma_X^2}\right)$$

(sample ≡ population)

[3.28'] $$\theta = \text{plim}(d) = \beta_{YZ\cdot X}\beta_{XZ}\sigma_Z^2\left(\frac{\sigma_X^2 - \sigma_{\bar{X}}^2}{\sigma_{\bar{X}}^2\sigma_X^2}\right) ,$$

(sample ≠ population, $N \to \infty$)

$$[3.29] \qquad \theta = E(d) = \beta_{YZ \cdot X}\beta_{XZ}E\left[(\Sigma z^2)\left(\frac{\Sigma x^2 - \Sigma \bar{x}^2}{\Sigma \bar{x}^2 \Sigma x^2}\right)\right]$$

$$+ \beta_{YZ \cdot X}E\left[(\Sigma zv)\left(\frac{\Sigma x^2 - \Sigma \bar{x}^2}{\Sigma \bar{x}^2 \Sigma x^2}\right)\right]$$

(sample ≠ population)

Section V.A has already discussed the conditions under which bias occurs. In Chapter 6 we shall examine the bias of the slope estimates from several grouping variables by substituting empirical estimates of the model parameters into equation [3.19'].

At this point, however, we can get some idea about the bias for Category I grouping by examining the bias in estimated coefficients when the variables from the ungrouped model have been standardized before grouping.[8] Assume that the X_{ij}, Y_{ij}, and Z_{ij} are standardized. Let m groups of equal size n be formed on discrete values of Z so that $Z_{ij} = \bar{Z}_{i\cdot}$. Under these conditions,

(1) $\sigma_Z^2 = \sigma_{\bar{Z}}^2 = 1$,

[8]The practice of standardizing the variables before grouping serves two useful purposes. First, it places the regression coefficients on a uniform scale (0 to 1.0). Second, the coefficient from the regression of Y on X when both have unit variance equals the correlation between Y and X. This suggests a potentially useful way to estimate zero-order correlation coefficients from grouped data is to regress $\bar{Y}$ on $\bar{X}$ when the ungrouped variables have been standardized.

$$(2) \quad \sigma_v^2 = \sigma_X^2 - \beta_{XZ}^2\sigma_Z^2 = 1 - \beta_{XZ}^2 \quad ,$$

$$(3) \quad \sigma_{\bar{v}}^2 = \sigma_v^2/n \quad ,$$

and

$$(4) \quad \sigma_{\bar{X}}^2 = \beta_{XZ}^2\sigma_{\bar{Z}}^2 + \sigma_{\bar{v}}^2 = \beta_{XZ}^2 + (1 - \beta_{XZ}^2)/n$$

$$= [(n-1)\beta_{XZ}^2 + 1]/n \quad ,$$

where n is the number of observations per group (held constant over groups).

After substituting (1) and (4) in [3.19'], we obtain

$$[3.31] \qquad \theta^* = E(d^*) = \beta_{YZ\cdot X}\beta_{XZ}\left[\frac{(n-1)(1-\beta_{XZ}^2)}{(n-1)\beta_{XZ}^2 + 1}\right] \quad ,$$

where d^* denotes the discrepancy from estimating the regression coefficient for standardized observations from grouped data.

At this point we consider how the discrepancy varies according to the relations of Z to X and Y and according to the number of groups formed. To do this we assume that there is a pool of grouping variables, Z's , which have been standardized and have varying relations to X and Y (potentially different $\beta_{YZ\cdot X}$ and β_{XZ}). For simplicity we let the number of groups formed by a given Z vary according to the chosen grouping variable, but we assume that equal size groups are formed.

In Figure 3.2, bias, θ^* , is plotted against β_{XZ} with $\beta_{YZ\cdot X}$ fixed at .1 for selected values of n , where $N = nm$ is held

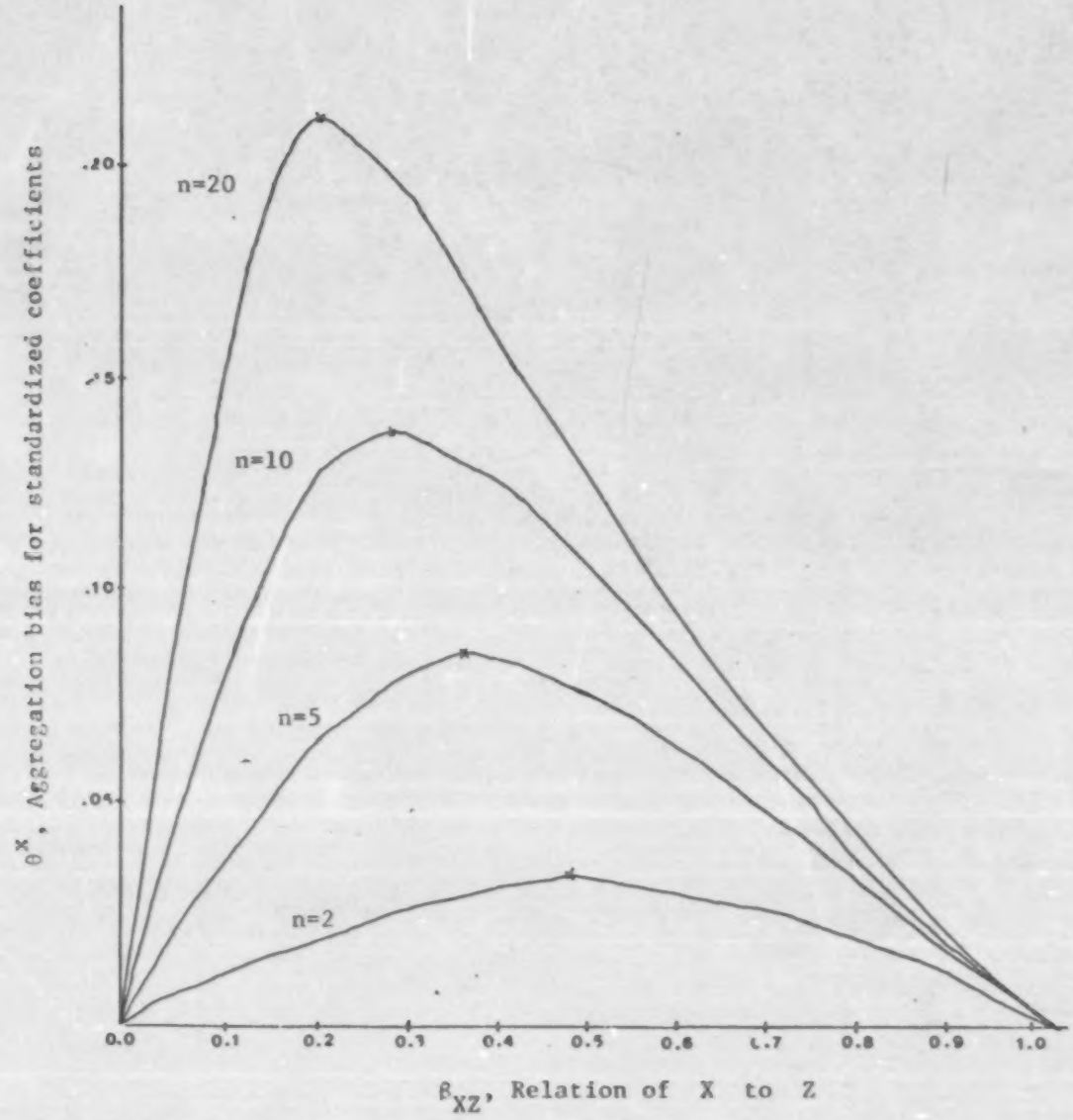


Figure 3.2. Aggregation bias θ^* (as defined by [3.31]) as a function of standardized β_{XZ} and group size n (with $\beta_{YZ \cdot X}$ fixed at .1).

constant. A comparable family of curves can be generated for any value of $\beta_{YZ\cdot X}$. The curves are roughly symmetrical for small n and become highly positively skewed for large n. This is as expected since the groupings become coarser and less representative of the ungrouped observations as n gets larger, for any set of fixed relations between Z and X and $Y\cdot X$.

Table 3.4 indicates the bias θ^* for several values of standardized $\beta_{YZ\cdot X}$, standardized β_{XZ}, and n. An examination of the tabled values leads to the following conclusions:

1) For any fixed values of $\beta_{YZ\cdot X}$ and β_{XZ}, bias increases with n (except $\beta_{YZ\cdot X} = 0$ or $\beta_{XZ} = 1$).

2) For fixed β_{XZ} (not 0 or 1) and n, bias increases with $\beta_{YZ\cdot X}$.

3) For fixed $\beta_{YZ\cdot X}$ (not 0 or 1) and n, bias first increases and then decreases as β_{XZ} goes from 0 to 1.

Minimizing the direct relation of Z to $Y\cdot X$ and maximizing the direct relation of Z to X is the safest way to reduce small bias. θ^* approaches its maximum rapidly even for small values of n. Large n is less damaging when β_{XZ} is large and $\beta_{YZ\cdot X}$ is small, though the necessary value of β_{XZ} increases rapidly with $\beta_{YZ\cdot X}$. For $n = 500$ and $\beta_{YZ\cdot X} = .1$, β_{XZ} must be greater than .60 to have bias less than .1. For $n = 500$ and $\beta_{YZ\cdot X} = .2$, β_{XZ} must be greater than .78 to achieve the same results.

The bias from Category I grouping can exceed 1 with large n and $\beta_{YZ\cdot X} > \beta_{XZ}$. This should be a further warning against choosing a grouping variable strongly related to $Y\cdot X$ and against concentrating

Table 3.4 Bias θ^* in estimating standardized regression coefficient β_{YX} from grouped data as a function of group size, standardized $\beta_{YZ \cdot X}$ and standardized β_{XZ}.

Group Size n	θ^* - Magnitude of the Bias[a]								
	$\beta_{YZ \cdot X}$ =	.2			.5			.8	
	β_{XZ} =	.2	.5	.8	.2	.5	.8	.2	.5	.8
2		.037	.060	.035	.093	.150	.088	.148	.240	.140
4		.103	.129	.059	.258	.323	.148	.412	.516	.236
5		.132	.150	.065	.330	.375	.163	.528	.600	.260
11		.274	.214	.078	.685	.535	.195	1.096	.856	.312
20		.415	.248	.083	1.038	.620	.208	1.660	.992	.332
50		.636	.277	.087	1.590	.693	.218	2.544	1.108	.348
100		.766	.288	.089	1.915	.720	.223	3.064	1.152	.356
500		.914	.298	.090	2.285	.745	.225	3.656	1.192	.360

$$^{a}\theta^* = \beta_{YZ \cdot X}\beta_{XZ}\left[\frac{(n-1)(1-\beta_{XZ}^2)}{(n-1)\beta_{XZ}^2 + 1}\right]$$

observations in a few large groups. On the other hand, the relatively small bias expected with small $\beta_{YZ\cdot X}$ offers some hope for reasonable estimates from data grouped by Category I variable.

2. Category II -- Z directly related to Y·X but not to X ($\beta_{YZ\cdot X} \neq 0$, $\beta_{XZ} = 0$) .

Category II contains grouping variables Z which are related to Y ($\beta_{YZ\cdot X} \neq 0$) and are not related to X ($\beta_{XZ} = 0$) . Since $\beta_{XZ} = 0$, $\Sigma xz = 0$, and regardless of whether the sample equals the population, the bias $\theta = E(d) = 0$, <u>as long as</u> $\Sigma \bar{x}^2 \neq 0$.

Thus estimates derived from data grouped by a Category II variable are unbiased unless there is no between-group variation in X . This conclusion is not surprising. When Z is a Category II variable, we are considering the standard model of equation [3.1] where the "other" determiners represented by u have been divided into two parts (Z and w), both independent of X . Unbiased estimates are expected under these conditions.

It is possible to have no between-group variation in X for a Category II variable. This occurs when the grouping variable lies in the X,Y plane; i.e., if $R^2_{Z\cdot X,Y} = 1$. In this case $\Sigma \bar{x}^2 = 0$ and since $\beta_{XZ} = 0$, the bias from grouping is indeterminate as can be seen by substitution into [3.19']:

$$\theta = \beta_{YZ\cdot X}\beta_{XZ}\sigma_Z^2\left(\frac{\sigma_X^2 - \sigma_{\bar{X}}^2}{\sigma_X^2\sigma_{\bar{X}}^2}\right)$$

$$= \beta_{YZ\cdot X}(0)\sigma_Z^2\left(\frac{\sigma_X^2 - (0)}{\sigma_X^2(0)}\right) .$$

There is no simple way to consider further the magnitude of the bias. There is some evidence based on simulation studies that bias estimates fluctuate wildly in this special case.

Category II variables are hard to find. None of the more than 200 pairs of parameter estimates, $\beta_{YZ\cdot X}$ and β_{XZ}, from the empirical data discussed in Chapter 6 satisfactorily meet the conditions for Category II grouping. Such variables could be constructed by orthogonalization, but other categories of variables yield unbiased estimators with greater efficiency. Henceforth, Category II will receive little attention.

3. Category III -- Z directly related to X but not to Y·X ($\beta_{YZ\cdot X} = 0$, $\beta_{XZ} \neq 0$) .

Category III includes variables which are related to Y only through X . Systematic grouping on the independent variable falls in this category. A Category III variable may be an explicit ordered function of X such as the decile rank of X , and if so, the within-group distributions of X do not overlap. It is also possible that a Z from Category III involves some random component (v) which allows the within-group distributions of X to overlap. The presence or absence of overlap is irrelevant in the determination of bias, but it can affect efficiency.

Since $\beta_{YZ\cdot X} = 0$ for Category III, equations [3.14a] and (3.17a] reduce to

$$Y = \alpha + \beta_{YX}X + w$$

and

$$\bar{Y} = \alpha + \beta_{YX}\bar{X} + \bar{w}$$

These equations are the same as [3.1] though the disturbance terms have been relabeled. Thus for Category III grouping, the standard model and our modified structure with the grouping variable incorporated are the same, and estimate the same β_{YX} .

From equations [3.19'], [3.28'], and [3.29'], it follows that when Z is a Category III variable,

$$E(B_{\overline{YX}}) = E(b_{YX}) = \beta_{YX} \quad ,$$

$$\text{plim}(B_{\overline{YX}}) = \text{plim}(b_{YX}) = \beta_{YX} \quad ,$$

and

$$\theta = E(d) = 0 \quad .$$

Thus the least-squares estimators of β_{YX} from data grouped on a variable Z which is related to X but not to Y·X are unbiased for any value of β_{YX}.

The bias and efficiency resulting from grouping by a function of X (Category III grouping) have been studied extensively, the most prominent being the Prais and Aitchinson study (1954). (Most variables systematically related to X do not strictly satisfy the condition $\beta_{YZ \cdot X} = 0$ and thus exhibit some minimal bias.) Our conclusions confirm those of earlier writers that Category III variables yield the best estimates under a very general set of analysis situations. The estimates are always unbiased and can be highly efficient (see Section V.C.2). If such variables do exist in a study, the remaining decision should focus on choice among Category III variables, and, once a variable is chosen, on the definition of the classes. These problems are considered in Chapter 4 under the heading of within-category factors.

4. Category IV -- Z not linearly related to X or Y·X .

($\beta_{YZ \cdot X} = 0$, $\beta_{XZ} = 0$) .

Category IV contains all variables which have no linear relation to either X or Y . A Category IV variable can be generated by assigning numbers randomly to individuals, such as a student ID. Category IV grouping, alternatively called random grouping, generates random groups of (X,Y) observations.

When $\beta_{YZ \cdot X} = 0$ and $\beta_{XZ} = 0$, it follows that

$$E(B_{\bar{Y}\bar{X}}) = E(b_{YX}) = \beta_{YX} \quad ,$$

and

$$\text{plim}(B_{\bar{Y}\bar{X}}) = \text{plim}(b_{YX}) = \beta_{YX} \quad .$$

Hence,

$$\theta = E(d) = 0$$

for any Category IV variable, and $B_{\bar{Y}\bar{X}}$ is an unbiased estimator of β_{YX}.

The interpretation of this result is straightforward. Estimating β_{YX} from the means of m randomly formed groups is statistically equivalent to estimating β_{YX} from a sample of size m drawn randomly from the N observations or from the m stratum means where the strata have been randomly formed (Hansen, Hurwitz, and Madow, 1953). In either case, the random process does not alter any pre-existing relations among the variables. All variances and covariances among variables decrease in proportion to the number of observations in a group for fixed group size n for Category IV grouping. This proportionate reduction in magnitude does not alter the estimate of the regression coefficient.

Category IV variables are not the best choice for grouping when efficient estimates are desired because of the difficulty of obtaining an adequate number of groups to overcome the marked efficiency reduction (see Section V.C.1.). In certain instances, however, Category IV variables may be the only recourse for the investigator who has limited information about other ways of forming groups.

C. Efficiency Considerations

Equation [3.12] defines efficiency. Below we evaluate the efficiency for each category of grouping variables.

1. Category IV

For Category IV variables, since $\beta_{YZ\cdot X} = 0$ and $\beta_{XZ} = 0$, equation [3.11'] becomes

$$\mathrm{Eff}(b_{YX}, B_{\bar{Y}\bar{X}}) = \frac{E[\frac{1}{SS_T(X)}]}{E[\frac{1}{SS_B(X)}]} \quad .$$

Several investigators have already provided simplified expressions for the efficiency of random grouping under the assumption that the X are fixed and given. An especially cogent derivation by Feige and Watts (1972) is presented below, using our terminology and notation.

Feige and Watts' derivation is based on the theory of sampling from a finite population. The set of N observations is regarded as a population. If the observations are assigned randomly to m groups of n_i in each group, many groupings are possible. The expected within-group sum of squares for the ith group is $SS_T(X)\ [(n_i-1)/(N-1)]$. Therefore, for Category IV grouping, the expectation of the total sum of squared deviations from the group means (the within-group sum of squares) is

$$E[SS_W(X)] = E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2\right]$$

$$= \sum_{i=1}^{m}\left[SS_T(X)\left(\frac{n_i - 1}{N - 1}\right)\right]$$

$$= SS_T(X)\left(\frac{N - m}{N - 1}\right) \quad .$$

From the formula for the decomposition of the total sum of squares, the expectation of the between-group sum of squares for Category IV grouping can be written as

$$E[SS_B(X)] = E[SS_T(X)] - E[SS_W(X)]$$

$$= E[SS_T(X)[- \frac{N-m}{N-1} E[SS_T(X)] \text{ (substituting from above)}$$

$$= \frac{m-1}{N-1} E[SS_T(X)] \quad .$$

If the X_{ij} are fixed and given,

$$E[SS_B(X)] = SS_B(X)$$

and

$$E[SS_T(X)] = SS_T(X) \quad .$$

Also, in Section III.D, we showed that if, in addition, $B_{\overline{YX}}$ is an unbiased estimator of β_{YX}, the efficiency of grouping is given by

$$Eff(b_{YX}, B_{\overline{YX}}) = \frac{SS_B(X)}{SS_T(X)} \quad .$$

Hence, by substitution for $SS_B(X)$, the efficiency of Category IV grouping when the X_{ij} are fixed and given is equal to

$$Eff(b_{YX}, B_{\overline{YX}}) = \frac{\frac{m-1}{N-1} SS_T(X)}{SS_T(X)}$$

$$= \frac{m-1}{N-1} \quad .$$

At best -- when there are $m = (N/2)$ groups of two observations each -- the efficiency of Category IV grouping is only about .5, under the assumption that the X_{ij} are fixed and given. However, the efficiency of random grouping provides a standard to which we can compare the efficiency of grouping in a systematic manner. Only those estimates with efficiency greater than $(m-1)/(N-1)$ offer an improvement over random grouping.

2. Category III

Category III grouping can produce small values of $E[\frac{1}{SS_B(X)}]$

because such groupings presumably assign observations to groups in part on the basis of their X values. Since maximization of the between-group sum of squares is a criterion for minimizing information loss through grouping, we expect Category III grouping to yield relatively efficient estimates.

Prais and Aitchinson (1954) and Cramer (1964) have examined the efficiency of grouping on the independent variable under the assumption that the X_{ij} are fixed and given. While they discussed grouping in a seemingly general way, their methods and conclusions are applicable to our Category III variables. Prais and Aitchinson presented a particularly illuminating example. They let X take on the mn equally-spaced values, $X_{ij} = 1, \ldots, mn$. Then adjacent observations were grouped into m groups of equal size and each value of X_{ij} was replaced by its group mean $\bar{X}_{i\cdot}$. Therefore, $\bar{X}_{i\cdot}$ takes on the values $[(2i-1)n + 1/2]$ where $i = 1, \ldots, m$.

$$SS_T(X) = \frac{mn(m^2n^2-1)}{12}$$

for the ungrouped observations and

$$SS_B(X) = \frac{mn^3(m^2-1)}{12}$$

for the grouped values. Whence,

$$\begin{aligned} \mathrm{Eff}(b_{YX}, B_{\bar{Y}\bar{X}}) &= \frac{m^2n^2-n^2}{m^2n^2-1} \\ &= 1 - \frac{n^2-1}{m^2n^2-1} \\ &> 1 - \frac{1}{m^2} \end{aligned}$$

In this special case of Category III grouping with fixed X_{ij}, then,

$$1 - \frac{1}{m^2} < \text{Eff}(b_{YX}, B_{\bar{Y}\bar{X}}) < 1 \quad .$$

[For $n < m$, $\text{Eff}(b_{YX}, B_{\bar{Y}\bar{X}})$ is also greater than $(1 - \frac{1}{n^2})$.]

Thus the lower bound of the efficiency of grouping related to X under these conditions depends only on m, the number of groups formed. Unfortunately, the distribution of observations seldom approaches this special case. The conditions under which the efficiency of other Category III variables approach this case are discussed in Chapter 4.

3. Category II

In Category II grouping, Z and X are stochastically independent. Category II variables share this property ($\beta_{XZ} = 0$) with Category IV variables. Since the efficiency of grouping is a function only of the variation of $\bar{X}$ and X when the estimators are unbiased, the efficiency of Category II grouping is the same as for Category IV grouping. That is, when $SS_B(X) \neq 0$, we expect Category II grouping also to have efficiency on the order of $(m-1)/(N-1)$, the ratio of the number of groups to the number of observations. It appears that neither Category II nor Category IV grouping yields estimators that approach the efficiency of the estimators from Category III.

4. Category I

When Z is a Category I variable, both bias and variance of $B_{\bar{Y}\bar{X}}$ affect the efficiency of estimation. Thus equation [3.12] defines the efficiency of grouping for this category of variables. In its simplest form, the efficiency of Category I grouping is given by

$$[3.12] \qquad \text{Eff}(b_{YX}, B_{\bar{Y}\bar{X}}) = \frac{V(b_{YX})}{V(B_{\bar{Y}\bar{X}}) + \theta^2} \quad .$$

If we again assume that the X_{ij} are fixed and given,

$$V(b_{YX}) = \frac{\sigma_w^2}{SS_w(X)} ,$$

$$V(B_{\overline{YX}}) = \frac{\sigma_w^2}{SS_B(X)}$$

and thus [3.12] can be written as

$$[3.32] \quad Eff(b_{YX}, B_{\overline{YX}}) = \frac{\sigma_w^2/SS_T(X)}{\left[\sigma^2/SS_B(X)\right] + \theta^2}$$

$$= \left(\frac{\sigma_w^2}{\sigma_w^2 + \theta^2 SS_B(X)}\right)\left(\frac{SS_B(X)}{SS_T(X)}\right)$$

$$< \eta_X^2 .$$

That is, the correlation ratio is an upper bound for the efficiency of Category I grouping when the X_{ij} are fixed and given.

One implication of the above is that grouping by a Category I variable is never more efficient than grouping by a Category III variable with comparable $SS_B(X)$. But since grouping randomly provides a lower bound for the efficiency of grouping when $B_{\overline{YX}}$ is an unbiased estimator of β_{YX}, Category I grouping can be more efficient than random grouping when θ is small.

For example, assume that 50 equal-size groups of 20 are formed. Let $\beta_{YX\cdot Z} = .5$, $\beta_{YZ\cdot X} = .2$, and $\beta_{XZ} = .8$. Also, assume that $\sigma_X^2 = \sigma_Y^2 = \sigma_Z^2 = \sigma_{\overline{Z}}^2 = 1$. Then, after solving for w in [3.14a] and remembering that w is unrelated to X and Z, we have

$$\sigma_w^2 = \sigma_Y^2 - \beta_{YX\cdot Z}^2\sigma_X^2 - \beta_{YZ\cdot X}^2\sigma_Z^2 - 2\beta_{YX\cdot Z}\beta_{YZ\cdot X}\sigma_{XZ}$$

$$= 1 - (.5)^2 - (.2)^2 - 2(.5)(.2)(.8)$$

$$= .55 \quad .$$

Also, from formula (4) on page 73,

$$\sigma^2_{\bar{X}} = [(n-1)\beta^2_{XZ} + 1]/n$$

$$= [(19)(.8) + 1]/20$$

$$= .678 \quad .$$

Hence

$$\eta^2_X = .658$$

and

$$SS_B(X) = (999)(.658) = 657.34 \quad .$$

From Table 3.4 we get the predicted bias for our chosen values of $\beta_{YZ\cdot X}(.2)$, $\beta_{XZ}(.8)$ and $n(20)$: $\theta = .083$ $(\theta^2 = .007)$.

Substituting the above in [3.32], we get

$$Eff(b_{YX}, B_{\bar{Y}\bar{X}}) = \frac{.55}{.55 + (.007)(657.34)} \quad (6.58)$$

$$= (.107)(.658)$$

$$= .070 \quad .$$

In comparison, the estimated efficiency from forming 50 groups of size 20 randomly is

$$Eff(b_{YX}, B_{\bar{Y}\bar{X}}) = \frac{m - 1}{N - 1}$$

$$= \frac{49}{999}$$

$$= .049 \quad .$$

Thus it is possible to improve efficiency relative to random grouping by grouping on a variable which yields small bias but is strongly related to X . By similar reasoning, we conclude that in certain cases, Category I grouping can yield more efficient estimators than Category II

grouping also.

VI. The Taxonomy as A Guide for Investigation

The main implication from the above discussion is that the investigator should consider the relations of the alternative grouping variables to the study variables before collecting his data, using such prior knowledge as is available. This will enable him to collect information on only those grouping variables that yield estimates having the desired properties.

If the investigator demands an unbiased estimate of β_{YX}, then, under the assumptions of the model, variables from Categories II, III, and IV can be satisfactory. While Category IV variables can always be created, they are relatively inefficient. Category III variables can be highly efficient, yielding large values of $SS_B(X)$. The efficiency of Category II grouping is no better than that of Category IV grouping because observations are assigned to groups essentially randomly with respect to X. Category III variables are clearly the best choice for data aggregation.

Category I variables yield biased estimates though the bias can be small with large β_{XZ} and small $\beta_{YZ \cdot X}$. Category I estimators are less efficient than Category III estimators but can be more efficient than those from Category II or Category IV grouping. If small bias is tolerable and Category III variables are hard to find, Category I grouping may be advisable.

Most of the discussion has assumed that an investigator has the original observations and can choose his own grouping procedure. Data can be available in aggregated form only, however; e.g., when individual data have been aggregated for economy of storage or for confiden-

tiality. The grouping variables that generally appear under these circumstances are geographic variables such as "state" and "census tract", and system delimiters such as "school" and "classroom". These grouping variables are generally related to X and $Y \cdot X$ and hence are Category I variables. Regression estimates determined under these conditions should be interpreted cautiously.

CHAPTER 4

ADDITIONAL CONSIDERATIONS IN THE SINGLE-REGRESSOR CASE

Until now, the discussion has concentrated on the effects of the linear relations of the grouping characteristic to the main variables on the precision of estimation from grouped observations. Other properties of the grouping characteristic -- the number and size of the groups it generates, its distribution, its scale of measurement -- need to be examined. Here we describe how these within-variable properties or factors affect the "utility" of a possible grouping variable.

Under the heading of properties of the distribution of observations, we consider the coarseness of grouping, the distribution of observations among the groups, and the distribution of the values of the independent variables both within and among the groups. These factors can often be manipulated by the investigator to improve estimation procedures.

Then, under the heading of scale of measurement, we discuss several methods for handling nominal[1] characteristics, such as school census tract. Such characteristics are of vital concern in recent educational investigations (see Averch et al., 1972). We consider in detail two related approaches to the problem. One approach [suggested by Wiley (personal communication)] provides a general scheme for classifying grouping variables on the basis of the scale (interval or nominal) and the type of variable (fixed or random). The other approach employs dummy coding to generate dichotomous variables to represent the grouping

[1] The discussion also applies to ordinal characteristics which are not transformed and treated as interval.

characteristic. The investigator then examines how properties of the dummy variables affect the proportion of variation accounted for in the model. This discussion relies less on formal mathematics than the preceding chapter. However, our exposition is tied conceptually to historical developments in the mathematics of scales of measurement and distribution. For our part, we are attempting to elaborate how the properties create distortions in empirical investigations of aggregated data.

I. Distributional Factors

In Chapter 3 we indicated that alternative grouping variables can be generated from a single grouping characteristic. Each grouping variable provides a unique classification of the individual observations. Thus, if groups are formed on achievement quartiles, the "grouping variable" is four-valued. There is one for each quartile; the finer subdivision by percentiles, or by score points is ignored. How to subdivide the scale is often under the investigator's control. This is particularly true of characteristics that have quasi-continuous distributions, e.g., "age" and "test score". There may also be a choice in subdividing a nominal grouping variable. Thus, race can be subdivided into "Anglo" and "Non-Anglo" or into "Anglo", "Asian-American", "American Indian", and so on.

In this section we examine the within-variable factors that are affected by the manipulation of the class boundaries of a given grouping characteristic using as an example the variables parental income (X) and family expenditures on higher education (Y).

Suppose that educational background is taken as the basis for grouping. The investigator can choose the number of groups (classes)

"m" for educational background and the location of the class boundaries. Table 4.1 illustrates several possibilities for subdividing educational background. $Z_{(5)}$ is a five-group classification and $Z_{(10)}$ and $Z'_{(10)}$ are ten-group breakdowns. With fixed m the number of cases per group and the skewness of the distribution depends on the boundaries.

Since the classifications of educational background in Table 4.1 give different $SS_B(X)$, the efficiencies of the grouped estimators they generate also differ. We explore these factors systematically below.

A. Coarseness of Grouping

In Chapter 3 we found that the coarseness of grouping, by which we mean the number of groups formed (m) for a fixed number of observations (N), has important effects on both bias and efficiency of grouping. According to equation [3.31], bias is inversely related to m. In addition, the efficiency of grouping increases with the number of classes. This finding has been supported through analyses of empirical and hypothetical data by several investigators (Blalock, 1964; Cramer, 1964; Prais and Aitchinson, 1954).

The effect of m on efficiency has already been discussed in connection with random grouping. The present discussion extends the "coarseness" principle to the more general case where the grouping variable is nonrandom. In our example, either $Z_{(10)}$ or $Z'_{(10)}$ yields a more efficient estimate then $Z_{(5)}$. With non-zero β_{XZ}, the groups of $Z_{(10)}$ tend to be more homogeneous than those of $Z_{(5)}$. In other words, the within-group variation of income and educational background is smaller with the ten-group classification of educational background than with the five-group classification. This means that the corresponding between-group variation is larger with the total variation

Table 4.1. Alternative grouping variables based on the same grouping characteristic.

	Grouping Variables		
	$Z_{(5)}$	$Z_{(10)}$	$Z'_{(10)}$
Classes Describing Father's Education	0-6 Years	None	0-6 Years
	7-10 Years	1-2 Years	7-10 Years
	11-HS Diploma	3-4 Years	11-HS Diploma
	1-3 Yrs. Beyond HS	5-6 Years	1-2 Yrs. Beyond HS
	More than 3 Beyond HS	7-8 Years	3-4 Yrs. Beyond HS
		9-10 Years	Bachelor's Degree
		11-12 Years	Work Beyond Bachelor's
		13-14 Years	Master's Degree
		15-15 Years	Work Beyond Masters
		More than 16 Years	Degree Beyond Masters (PhD, MD, LLD, etc.)

held constant. So the correlation ratio η_X^2 of either $Z_{(10)}$ or $Z'_{(10)}$ is greater than that of $Z_{(5)}$ and the estimate more efficient.

Cramer's paper (1964, p. 241) provides a particularly illuminating analysis of this topic. He considers the case where the individual observations are ordered according to their X values and the sample range is divided into m equal intervals. The total sum of squares is partitioned into between-groups and within-groups sums of squares, and the components are divided by the total. After rearranging terms, Cramer arrives at the efficiency equation:

$$[4.1] \qquad \frac{SS_B(X)}{SS_T(X)} = 1 - \frac{SS_W(X)}{SS_T(X)}$$

where $SS_W(X)$ is the pooled within-group sum of squares of the X_{ij}.

Cramer then estimates $SS_T(X)$ and $SS_W(X)$. For the sample of original observations,

$$SS_T(X) = N\sigma_X^2 \quad ,$$

where σ_X^2 is the population variance.

For his grouping method, the width of all class intervals is uniform and equals

$$\left[\frac{\text{range}(X)}{m}\right]\sigma_X^2 \quad ,$$

where the sample range of X is expressed in terms of the population standard error. Cramer then states that if the sample X_{ij} are uniformly distributed within each class, the within-group variance, $WV(X_{ij})$ of each class is

$$WV(X_{ij}) \cong \frac{1}{12}\left[\frac{\text{range}(X)}{m}\right]^2\sigma_X^2 \quad .$$

So the pooled within-class variation is approximated by

$$[4.2] \qquad SS_W(X) \cong \frac{N}{12}\left[\frac{\text{range}(X)}{m}\right]^2\sigma_X^2 \quad .$$

By substituting [4.2] into [4.1], we obtain the approximation

$$[4.3] \qquad \frac{SS_B(X)}{SS_T(X)} \approx 1 - \left[\frac{\text{range}(X)}{12m^2}\right]^2 .$$

Cramer points out that his approximation is justified for large N and relatively small m because it depends on the replacement of random variables by their expected values. He also emphasizes that his estimate of the within-groups variation of X is an overestimate when the actual distribution of X within class is a strip from the normal distribution, and not a rectangle.

One can use values from the sampling distribution of range (X) to provide efficiency estimates of various combinations of m and N. From Cramer, the expected values of range (X) with the sample sizes 100, 200, 500, and 1,000 are 5.015, 5.492, 6.073, and 6.483, respectively. Table 4.2 includes the efficiency of grouping N observations into m equal-interval groups. The values are in agreement with a similar table by Cramer (1964, p. 244).

Efficiency appears to be very high except with very small m. Most investigators would happily use group means when the efficiency of the regression estimate from grouped data is high to reduce cost of data processing.

Cramer describes a fairly representative method of grouping in economic studies. Unfortunately, his findings do not apply to Category III grouping variables with unequal intervals nor do they apply to variables in other categories. Equal-interval grouping may not be appropriate in many educational investigations. Thus we cannot expect estimates as efficient as those depicted in Table 4.2.

Table 4.2. Efficiency of alternative ways of grouping on the same characteristic as a function of sample size and number of groups.

Sample Size N	E[range(X)]*	Efficiency ($SS(\bar{X})/SS(X)$) Number of Groups					
		m=2	m=4	m=5	m=10	m=20	m=25
100	5.015	0.476	0.869	0.916	0.979	0.995	0.997
200	5.492	0.372	0.843	0.899	0.974	0.994	0.996
500	6.073	0.232	0.808	0.877	0.969	0.992	0.995
1000	6.483	0.123	0.781	0.860	0.965	0.991	0.994

*See page 93.

B. Distribution of Observations Among the Groups

The distribution of observations among the groups is of concern only when there are some groups with very few observations and when the independent variable is imperfectly measured. In the former case, some group means are unstable, and their instability reduces the precision of the grouped estimate.

A large number of observations per group are needed to cancel out the effects of random errors of measurement on the independent variable (Blalock, Carter, and Wells, 1971). In the example above, this can mean that $Z_{(5)}$ is better for grouping than $Z_{(10)}$, depending on the within-group distribution of the income values and on the size of the errors.

It is not always easy to determine whether there are enough observations per group for adequate stability. Generally, grouping variables with large skewness coefficients yield imprecise estimates. However, with other variables, groups with few observations are scattered along the Z scale. With these variables, the investigator must rely on his understanding of the nature of the grouping characteristic and its relation to other study variables to avoid imprecise estimates.

C. Distribution of the Independent Variable Within and Among Groups

Though $m = 10$ for both $Z_{(10)}$ and $Z'_{(10)}$ in Table 4.1, the two classifications yield equally efficient estimators only when $V(\bar{X}|Z_{(10)}|) = V(\bar{X}|Z'_{(10)})$. The subdivisions of these two classifications are not likely to result in equal between-group variances and the pooled within-group variation in X for $Z_{(10)}$ and $Z'_{(10)}$ are undoubtedly different. Thus the within-group distributions of X and the overlap of these distributions are affected by the placement of the

class boundaries, and, in turn, affect the efficiencies of grouping.

Even without a joint distribution of income and educational background, it is possible to envision the properties of this distribution after classification. With $Z_{(10)}$ the mean incomes and income ranges are approximately the same for the "none" through "7-8 years" groups. Hence, the income distributions of the groups from $Z_{(10)}$ overlap a great deal. Individually, some of the groups contributed little to the between-groups variance. In fact, collapsing the five lowest groups into a single "0-8 years" group does not greatly change the between-groups variance. So $Z_{(10)}$ acts rather like, say, a $Z_{(6)}$.

$Z'_{(10)}$, on the other hand, has wide intervals at the lower end, a relatively uniform distribution of observations, and large variation in group means. It forms homogeneous income groups by adding groups at the upper end and collapsing similar (in income) groups at the lower end. We suspect that $Z'_{(10)}$ forms income groups which are more compact (smaller within-group variation) and more distinct (less overlap among groups) than those from $Z_{(10)}$. If so, this combination is sufficient to ensure that the between-group variance in income will be greater with $Z'_{(10)}$, and its grouped estimator more efficient.

In general, classifications which yield small within-group variation in the independent variable are preferred. This type of classification decreases the pooled within-group variation and thus increases between-group variation.

The effects of overlap of the within-group distributions of the independent variable operate similarly. As the overlap among distributions decreases, grouping more closely resembles direct stratification on X, which is optimally efficient.

D. Summary

The within-variable factors that affect estimation are interdependent and tend to constrain each other. Insofar as finer breakdowns increase the relation between the grouping variable and the independent variable, information loss declines and precision increases. If the characteristic is judiciously chosen, The investigator can quickly arrive at a grouping which balances the competing factors and yields estimates which suit his purposes.

II. Scales of Measurement -- Nominal Grouping Characteristics

So far, we have treated the grouping characteristic as if it has at least an interval scale and thus has specifiable linear relations with the dependent and independent variables. The next step is to consider grouping characteristics that have nominal scales.

Sound procedures for predicting the effects of a nominal grouping characteristic are urgently needed in educational research. Cross-level inferences from aggregate sampling units such as schools occur frequently; careful examination of the consequences is needed. Unfortunately, the sociological methods developed to date are often complex, and some apply primarily to relations among unordered variables (Goodman, 1959; Iversen, 1973).

Our approach is to try to fit structural-equation methods to this case. We shall incorporate the nominal grouping characteristics into the model as we incorporated ordered characteristics. Two schemes for incorporating the nominal grouping characteristic are discussed below. Wiley (personal communication) actually offers a new conceptual scheme for analyzing the grouping process. The other approach, the creation of multiple dichotomies to represent the nominal characteristic, adapts a

familiar econometric technique.

A. Categorization by Scale and Type of Variable

To this point we have considered only the manifest relations of grouping variables to the other study variables. We have not attempted to describe the latent forces that underly the grouping of observations. When the manifest grouping characteristic has a nominal scale, a more careful examination of the classification process may prove useful. Classification procedures such as latent structure analysis have been discussed in this context. We consider here the implications of a procedure suggested by Wiley for aggregation problems.

1. The Classification Matrix

Wiley's scheme for classifying grouping variables is a variation of the model represented by the structural equations [3.14a] and [3.14b] and by the path diagrams in Figure 3.1. Additionally, however, (1) each Z is now said to be either "fixed" or a "random" variable, and (2) attention is now paid to whether it has either a nominal or interval scale.

Before, a grouping variable was spoken of as random if the individual observations were randomly allotted to groups. Here, Z is considered a random variable if the groups of Z are randomly sampled from some broader population. Z thus operates like a random factor in the analysis of variance as opposed to a fixed factor. Randomness is a property of the selection of groups, not of the assignment of observations to the groups.

To clarify Wiley's scheme, consider the following hypothetical data set. Suppose that data on the following grouping variables were collected in an international study of the relation of home environment to mathematics achievement: the sex of students, the nation, the classroom, the school, the school size, student mathematical aptitude,

and the salary of the student's math teacher.

We can classify each variable within a scale × type-of-variable grid. The nominal vs. interval dichotomy is relatively straightforward. In their present form, school size, math aptitude, and teacher salary are the variables with interval scales.

Classification by types of variable requires more thought. It is likely that the classrooms in the study are important only as "representative" of similar entities. Therefore, we can treat the classrooms as random samples from some larger population of classrooms.

Examining each grouping variable in the same manner leads to classification matrix A:

Matrix A

	FIXED	RANDOM
NOMINAL	Sex	Classroom Nation School
INTERVAL	School Size Math Aptitude Teacher Salary	

2. Manifest vs. Latent Grouping

Wiley argues that, in general, grouping characteristics like school and classroom are surrogates for some unmeasured variables which have interval scales. (Without loss of generality, we assume there is only one unmeasured variable.) In other words, there exists some underlying interval variable Z^{∞} which determines group membership when observations are manifestly grouped by a nominal variable Z^{+}. In our present example, this might mean that nation is really a proxy for, say, national commitment to education. Then grouping by nation would approximate grouping by national commitment to education (as measured on an interval scale).

We can illustrate the interrelation of Z^+ and Z^∞ by incorporating both in the path diagram. This model is presented in Figure 4.1. When the data are grouped by Z^+, Figure 4.2 represents the aggregate path diagram corresponding to Figure 4.1.

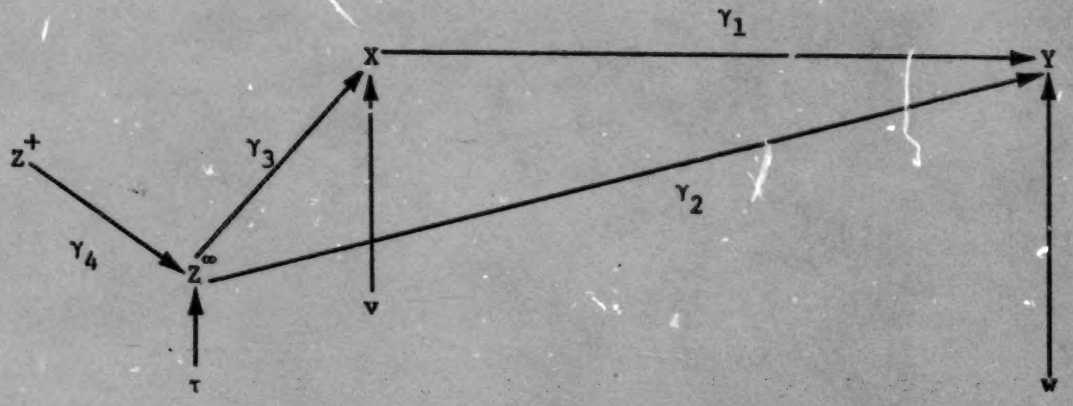
Given these path models, the investigation properly focuses on the conditions under which $\gamma_1 + \gamma_2\gamma_3 = \lambda_1 + \lambda_2\lambda_3$. The question to be answered is "Does grouping by Z^+ affect Z^∞ in a way that will change the relation of X to Y ?" If the answer is yes, then grouping by nationality yields biased estimates.

Z^∞ cannot be directly measured. It is a latent variable analogous to the latent traits of factor analytic models. However, values of Z^∞ can be estimated by $D(Z^+)$, a discriminant function describing the differences in the classes of Z^+ with respect to variables potentially influencing the grouping process.

In the example above, national commitment to education is the latent variable represented by nation. Substantial auxiliary information, such as per pupil expenditures (W_1), educational expenditure as a proportion of national GNP (W_2), and proportion of children enrolled in school at, say, age 15 (W_3) is needed to have a prayer that $D(Z^+)$ generates good estimates of Z values. The equation representing this relation would be

$$\text{National Commitment to Education} = D(\text{Nation}) + \delta$$
$$= \phi_1 W_1 + \phi_2 W_2 + \phi_3 W_3 + \delta$$

where the ϕ's are the variable weights in the discriminant function and δ represents unaccountable differences in national commitment. δ must approach zero if the grouped estimate is to be unbiased.



Z^+ -- manifest (or measured) grouping variable

Z^∞ -- latent (or unmeasured) grouping variable

Y -- dependent variable

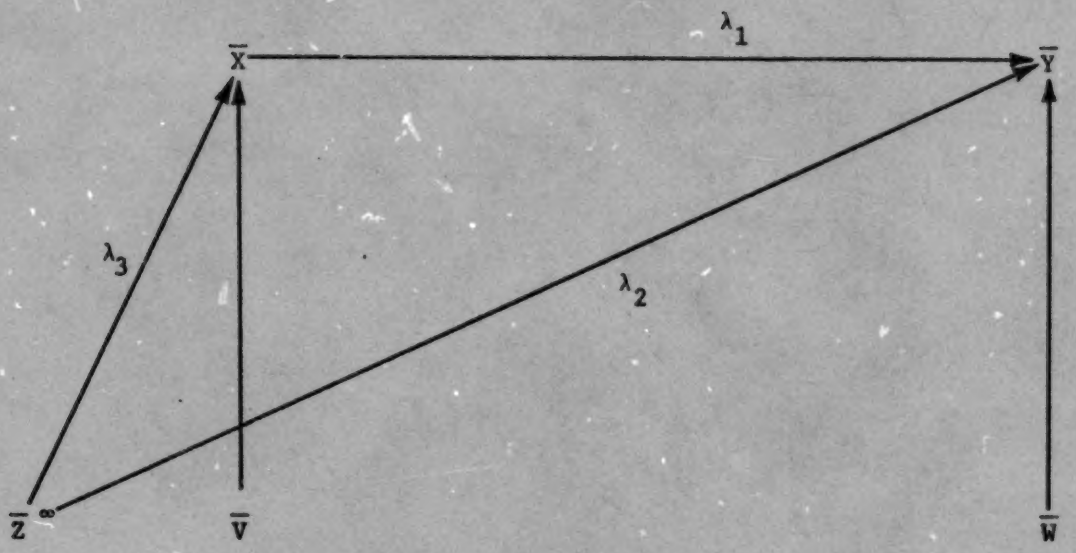
X -- independent variable

v, w -- disturbance terms for X and Y

γ_f -- structural parameter for relation designated by corresponding arrows

τ -- disturbance term for Z^∞

Figure 4.1. Path diagram incorporating both latent and manifest grouping variables.



$\bar{X}$ -- Aggregate independent variable (group means based on Z^+)

$\bar{Y}$ -- Aggregate dependent variable (group means based on Z^+)

$\bar{Z}^{\infty}$ -- Aggregate latent grouping variable

v, w -- Structural parameters for aggregate $\bar{X}$ and $\bar{Y}$

λ_f -- structural parameters for aggregate relations designated by corresponding arrows

Figure 4.2. Path diagram for aggregate data grouped by Z^+

This is a minimum condition for maintaining consistent relations among X , Y , and Z^{∞} at the individual and group levels. Otherwise, the influence of δ , which has an effect on X and Y independent of Z^{+} , will change between levels.

Returning to hypothetical data, we can conceivably estimate the Z^{∞}'s of the particular classrooms, schools, and nations. In fact, all the nominal variables can be handled in this way. If so then the new classification matrix would be

Matrix B

	FIXED	RANDOM
NOMINAL		
INTERVAL	School Size Math Aptitude Teacher Salary D(Sex)	D(Classroom) D(School) D(Nation)

3. Evaluation of the Wiley Classification Scheme

According to Wiley's scheme, we can always generate an interval grouping variable if enough information is available. The investigator cannot translate his knowledge of the underlying grouping variable into an ordered function without resorting to classification procedures of this sort.

At the same time, however, the search for an underlying grouping variable greatly complicates the procedure for choosing Z . Where before only estimates of $\beta_{YX \cdot Z^{+}}$, $\beta_{XZ^{+}}$, and $\beta_{YZ^{+} \cdot X}$ were needed, we must now find the underlying Z^{∞} . Besides, we still have to determine optimal class intervals (with respect to within-variable factors) for $D(Z^{+})$ after the variable has been generated.

The benefits from estimating $D(Z^{+})$ are derived mainly from the

uncovering of the inherent causal patterns among the grouping variables which affect the estimation of β_{YX}. If the investigator's efforts are directed toward "purity" in aggregation and more accurate specification of the model, Wiley's methods can be useful. It makes little sense, on the other hand, to estimate Z solely for the purpose of having an interval grouping variable.

The type-of-variable distinction raises serious questions about the process of grouping. If the classes of the grouping variable are fixed, then there is no change in the conceptualization of grouping effects. If, on the other hand, the classes are random, the original observations should be treated as a single or two-stage cluster sample rather than as a simple random sample for the purposes of grouping. In cluster sampling, the selected clusters (individual classrooms, for example) are a simple random sample from the population of clusters and sampling within the clusters is also random.

The distinction between cluster and simple random sampling apparently has not been made before in the context of grouping. The usual regression analyses start with the assumption that the data are a simple random sample. We do not find fault with this assumption for the ungrouped observations or for a fixed number of groups. The sampling properties of the data become an issue only after grouping. The question then arises as to whether the classes of Z can be considered a simple random sample since the classes become the units for analysis. An unbiased estimate is impossible if the groups themselves are a non-random sample, whether the units are the original observations or the weighted group means.

B. Dummy Coding

Economists generally employ dummy coding methods to incorporate

nominal characteristics in their models. This procedure is less complex than Wiley's and may prove fruitful for our purposes.

In applying dummy coding, we represent any nominal characteristic with m groups by $m-1$ (or m, depending on the computer program) dichotomous dummy variables in the basic structural equations. Equations [3.14a] and [3.14b], which incorporate the grouping characteristic become

$$[4.3a] \quad Y = \alpha + \beta_{YX \cdot Z_1,\ldots,Z_{m-1}} X + \beta_{YZ_1 \cdot X,Z_2,\ldots,Z_{m-1}} Z_1 + \ldots + \beta_{YZ_{m-1} \cdot X,Z_1,\ldots,Z_{m-2}} Z_{m-1} + v \,,$$

$$[4.3b] \quad X = \lambda + \beta_{XZ_1 \cdot Z_2,\ldots,Z_{m-1}} Z_1 + \ldots + \beta_{XZ_{m-1} \cdot Z_1,\ldots,Z_{m-2}} Z_{m-1} + w$$

where the Z_i, $i = 1, \ldots, m-1$; are the dichotomous variables representing group membership, and the $\beta_{YZ_i \cdot X,Z_1,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_{m-1}}$ and the $\beta_{XZ_i \cdot Z,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_{m-1}}$ are the structural parameters in the regressions with Y and X, respectively.

Then, if $R^2_{Y \cdot X}$ is the squared correlation coefficient of X and Y and $R^2_{Y \cdot X,Z_1,\ldots,Z_{m-1}}$ and $R^2_{X \cdot Z_1,\ldots,Z_{m-1}}$ are the squared multiple correlation coefficients from incorporating the dichotomous regressors based on Z, then the direct strength of the relation of Y to Z can be estimated from the square root of the variation accounted for by incorporating the dummy variables. That is, we estimate $\beta_{YZ \cdot X}$ from

$$\sqrt{R^2_{Y \cdot X,Z_1,\ldots,Z_{m-1}} - R^2_{Y \cdot X}} \,.$$

The relation of X to Z (β_{XZ}) can be estimated from

$$\sqrt{R^2_{X \cdot Z_1, \ldots, Z_{m-1}}} \quad .$$

This estimation procedure requires some justification. The reason for the use of the square root of the variation accounted for is to have units comparable to the standardized regression coefficients from incorporating interval grouping variables. The "additional variation accounted for" notion embodied in our suggested estimator $\beta_{YZ \cdot X}$ is an attempt to identify any relationship between Y and Z's that is masked in the simple linear model for ungrouped observations (Equation [3.1]). The estimator suggested for β_{XZ} provides an indication of the magnitude of the relation between X and Z's (η_X would also fulfill this function). In this way, we hope to make direct comparisons of the effects of nominal grouping characteristics with the effects of interval characteristics. For this reason alone, the dummy coding strategy provides a viable alternative to the classification procedures which necessitate a search for the latent causes of group membership.[2]

C. Summary

Neither Wiley's scheme nor the dummy coding approach yields perfect indices of the relations of a nominal Z to X and Y, but both warrant further consideration as alternatives to those previously proposed. They at least provide a starting point for refining the "structural equations" approach in the nominal case.

---

[2]Werts and Linn (1971) discuss the regression analysis for "compositional effects", which involves the incorporation of $\bar{X}_{i.}$, rather than Z in the simple model. Using $\bar{X}_{i.}$ instead of Z in the modified structure has the advantage of ensuring that the grouping mechanism is represented by an ordered variable, regardless of the scale of the grouping characteristic. However, with multiple regressors, this strategy can become cumbersome rapidly unless one incorporates, say, the values from the best linear discriminant function (discriminating among the Z values on the basis of a function of the X's). Still, the Werts-Linn method deserves more consideration than we have given it here.

CHAPTER 5

PRELIMINARY NOTES ON THE MULTIVARIATE CASE

Our findings on the effects of grouping in the bivariate case can be extended to the multivariate case. Problems caused by correlated regressors, however, can complicate the interpretation of grouping effects. These problems are considered below.

We begin by reviewing previous work on the multivariate case, considering papers by Prais and Aitchinson (1954), Haitovsky (1966; 1973), and Feige and Watts (1972). To simplify our own developments, we analyze the three-variable case where Y is regressed on just two independent variables, X and W. The grouping variable Z enters as a fourth variable. The parameters to be estimated are the regression coefficients $\beta_{YX \cdot W}$ and $\beta_{YW \cdot X}$. The conclusions are generalizable to any number of regressors.

The earlier taxonomy is expanded to consider the interrelation of Y, X, W, and Z for a specific causal ordering of X and W. This taxonomy is used to investigate the bias in estimating the regression coefficients.

I. Previous Work on the Multivariate Case

Whereas univariate prediction with grouped data has been considered by persons from several social science disciplines, the treatment of grouping effects with multiple predictors has remained purely in the domain of the econometricians. Prais and Aitchinson (1954) seemingly stood alone until Haitovsky (1966) suggested that grouping can indeed cause bias in the multivariate case. Feige and Watts (1972) --

apparently unfamiliar with Haitovsky's work -- raised much the same question. Below we attempt to reconcile the conclusions of Prais-Aitchinson, Haitovsky, and Feige-Watts.

A. Transformation by a Grouping Matrix -- Prais and Aitchinson

Prais and Aitchinson (1954) derived formulas for grouped estimation in the multivariate case. They employed matrix notation throughout.

Consider the usual postulated model for multiple linear regression:

[5.1] $$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{u} \quad ,$$

where $\underset{\sim}{Y}$, $\underset{\sim}{X}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{u}$ are matrices of orders $N \times 1$, $N \times k$, $k \times 1$, and $N \times 1$, respectively. We assume that the rank of $\underset{\sim}{X}$ is k , the number of regressors, where k is less than or equal to the number of persons N .

An estimate of $\underset{\sim}{\beta}$ can be found by the principle of least-squares (LS) . The assumptions in the multivariate case are analogous to those of the model (equation 3.1). They are as follows:

B1. The $\underset{\sim}{X}$ are fixed or else the $\underset{\sim}{X}$ are random variables with joint distribution independent of $\underset{\sim}{u}$.

B2. $E(\underset{\sim}{u}) = 0$.

B3. $V(\underset{\sim}{u}) = E(\underset{\sim}{u}\underset{\sim}{u}') = \left(\sigma_u^2\right)\underset{\sim}{M}_N$, where $\underset{\sim}{M}_N$ is a known matrix of order N .

B4. $\underset{\sim}{X}$ is of rank k .

The principle of least-squares provides an estimator of $\underset{\sim}{\beta}$ that minimizes the sum of squares of deviations of $\underset{\sim}{Y}$ and $\underset{\sim}{X}\underset{\sim}{\beta}$. This estimator is given by

[5.2] $$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y} \quad .$$

$\underset{\sim}{b}$ is an unbiased estimator of $\underset{\sim}{\beta}$. If the $\underset{\sim}{u}$ are normally distributed,

then $\underset{\sim}{b}$ is also a maximum likelihood estimator (MLE).

The covariance matrix of the vector $\underset{\sim}{b}$ is

$$[5.3] \qquad \text{cov}(\underset{\sim}{b}) = E[(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})'] = \sigma_u^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1} \quad ,$$

and the residuals $\underset{\sim}{e} = \underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{b}$ are linear functions of the disturbances $\underset{\sim}{u}$.

Prais and Aitchinson next introduced an m by N grouping matrix $\underset{\sim}{G}$ which maps the original observations into their appropriate groups and weights each group by the number of observations included. Thus $\underset{\sim}{G}$ is a weighting matrix in which the weights are determined by the number of observations in the various groups. The value in the ith row of $\underset{\sim}{G}$ is $1/m_i$ for persons belonging to group i and 0 for persons not in group i. For example, with five observations divided into two groups (m = 2) with the first, third, and fourth in the first group and the second and fifth in the other, the weighting matrix is

$$G = \begin{bmatrix} 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \end{bmatrix} .$$

Note that

$$\underset{\sim}{G}\underset{\sim}{G}' = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix}$$

and

$$(\underset{\sim}{G}\underset{\sim}{G}')^{-1} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$$

That is, the diagonal elements of the inverse of $\underset{\sim}{G}\underset{\sim}{G}'$ indicate the number of observations per group.

The regression model for the grouped data then found by premultiplying [5.1] by $\underset{\sim}{G}$ to get

$$\underset{\sim}{G}\underset{\sim}{Y} = \underset{\sim}{G}\underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{G}\underset{\sim}{u} \quad .$$

Since $\underset{\sim}{G}$ is a weighting matrix, it gives us means, i.e., $\underset{\sim}{GY} = \bar{\underset{\sim}{Y}}$, $\underset{\sim}{GX} = \bar{\underset{\sim}{X}}$, and $\underset{\sim}{Gu} = \bar{\underset{\sim}{u}}$. So we obtain

$$[5.4] \qquad \bar{\underset{\sim}{Y}} = \bar{\underset{\sim}{X}}\underset{\sim}{\beta} + \bar{\underset{\sim}{u}} \quad .$$

By assuming that the number of groups formed, m, is greater than or equal to the number of parameters estimated, it follows that $\underset{\sim}{GX}$ is of rank k. Consequently, the assumptions B1-B4 apply to the model [5.4] where

$$E(\bar{\underset{\sim}{u}}) = 0$$

and

$$V(\bar{\underset{\sim}{u}}) = \sigma_u^2 \underset{\sim}{GG}' \quad .$$

Under these conditions, the grouped estimator $\underset{\sim}{B}$ for $\underset{\sim}{\beta}$ is

$$[5.5] \qquad \underset{\sim}{B} = [\bar{\underset{\sim}{X}}'(\underset{\sim}{GG}')^{-1}\bar{\underset{\sim}{X}}]^{-1}\bar{\underset{\sim}{X}}'(\underset{\sim}{GG}')^{-1}\bar{\underset{\sim}{Y}}$$

$$= (\underset{\sim}{X}'\underset{\sim}{HX})^{-1}\underset{\sim}{X}'\underset{\sim}{HY}$$

where

$$\underset{\sim}{H} = \underset{\sim}{G}'(\underset{\sim}{GG}')^{-1}\underset{\sim}{G} \quad .$$

For $\underset{\sim}{X}$ fixed (or for $\underset{\sim}{X}$ random, because of assumption B1 and the fact that $E(\underset{\sim}{Y}) = \underset{\sim}{X\beta}$),

$$E(\underset{\sim}{B}) = [(\underset{\sim}{X}'\underset{\sim}{HX})^{-1}\underset{\sim}{X}'\underset{\sim}{H}]E(\underset{\sim}{Y})$$

$$= (\underset{\sim}{X}'\underset{\sim}{HX})^{-1}\underset{\sim}{X}'\underset{\sim}{HX\beta}$$

$$= \underset{\sim}{\beta} \quad .$$

The covariance matrix of $\underset{\sim}{B}$ is given by

$$[5.6] \qquad \text{cov}(\underset{\sim}{B}) = \sigma_u^2[\bar{\underset{\sim}{X}}'(\underset{\sim}{GG}')^{-1}\bar{\underset{\sim}{X}}]^{-1}$$

$$= \sigma_u^2(\underset{\sim}{X}'\underset{\sim}{HX})^{-1} \quad .$$

Prais and Aitchinson concluded that "whatever the method of grouping, the resulting estimators will always be unbiased" (1954, p. 1). But this contradicts the results of Chapter 3 for grouping in the single regressor case. The work by Haitovsky (1966; 1973) and by Feige and Watts (1972) and the new material in Sections II and III of this chapter identify limitations of their formulation which led them into difficulty.

Prais and Aitchinson also provided an overall measure of the efficiency of the method of grouping:

$$[5.7] \qquad \text{Eff}(\underset{\sim}{b},\underset{\sim}{B}) = \frac{\text{tr}(\underset{\sim}{X}'\underset{\sim}{H}\underset{\sim}{X})^{-1}}{\text{tr}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}} \quad ,$$

the ratio of the sum of the diagonal elements from the covariance matrix of $\underset{\sim}{b}$ to the corresponding sum from the covariance matrix of $\underset{\sim}{B}$. In the single-regressor case with $\underset{\sim}{X}$ fixed, their efficiency formula simplifies to become the ratio of the between-group sum of squares to the total sum of squares, the equivalent of Cramer's formula (see page 47). When there is no bias from grouping, this measure of efficiency is appropriate.

B. Estimates from Classification Data -- Haitovsky

Haitovsky (1966; 1973) called into question the conclusion of Prais and Aitchinson. He demonstrated problems that arise when the regressor data are in the form of one-way classification tables, with frequencies of the cross-classifications unknown. According to Haitovsky, grouping on one independent variable can lead to biased estimates of the multiple regression coefficients in this situation.

Haitovsky analyzed data from a study by Houthakker and Haldi (n.d) to illustrate his conclusions. In the Houthakker-Haldi study, automobile purchases (Y) were regressed on individual income (X) and

initial automobile inventory (W) . Haitovsky grouped observations on X and W separately as well as on the cross-classification of X and W . His estimates for $\beta_{YX \cdot W}$ and $\beta_{YW \cdot X}$ are presented in Table 5.1.

The estimates from the cross-classification were fairly accurate. The single-variable classifications yielded estimates with hugh standard errors. If 7 or 8 groups had been formed randomly, we would have expected the standard errors to be even larger.

Haitovsky failed to note an interesting trend in the data. When the observations were grouped by one regressor, say, X , its regression coefficient $\beta_{YX \cdot W}$ was better estimated, in terms of smaller bias and standard errors, than was the coefficient $\beta_{YW \cdot X}$ of the other regressor. That is, grouping on a regressor affected the estimate of its coefficient less than it did the estimate of the coefficients for other variables.

As Hannan (1972) put it, Haitovsky's paper showed that "in the multivariate model, grouping by some concrete criterion which approximates grouping systematically by a subset of the regressors ... can produce appreciable bias." (p. 33). Hannan also pointed out that the bias Haitovsky described is essentially specification bias. That is, bias arises through the failure to include all correlated regressor variables in the data analysis.

According to Haitovsky and Hannan, unbiased estimates are obtainable if the investigator groups on all regressor variables jointly. But with a large number of independent variables each having several classifications, grouping on all jointly is impractical. Unless other

Table 5.1. Estimates of regression coefficients and standard errors with alternative grouping methods from the Houthakker-Haldi study[a].

Grouping Method	Number of Groups (m)	$\hat{\beta}_{YX \cdot W}$	$\hat{\beta}_{YW \cdot X}$
Ungrouped data	1218	.758 (.1398)	-.178 (.0367)
Income(X)-x-Inventory (W)	56	.747 (.1203)	-.162 (.0323)
Income (X) only	7	.551 (1.6139)	.038 (1.9752)
Inventory (W) only	8	-.653 (2.5391)	-.093 (.1572)

[a]The numbers in parentheses are the estimated standard errors of the corresponding estimates.

evidence is forthcoming, it is easy to agree with Hannan's conclusion that the analyst must have a good deal of confidence in the substantive aspects of his model before concluding that any grouping procedure is "optimal".

C. Aggregating Data to Preserve Confidentiality -- Feige and Watts

Feige and Watts (1970; 1972) considered the feasibility of data aggregation as a means of preserving the confidentiality of data. They developed statistics for evaluating the loss of information from grouping in this context. One measure indicates the degree of divergence between estimates from grouped and ungrouped data, and the other indicates the loss of efficiency. Feige and Watts applied a variety of grouping procedures to a large data set and assessed the resulting parameter estimates.

According to Feige and Watts, differences between the ungrouped and grouped estimators may be composed of (i) specification bias, (ii) bias introduced by a grouping that is not independent of the disturbances, or (iii) sampling error induced by the loss of information in grouping. Their second source is most pertinent to our discussion since it suggests that even when the regressors and disturbances are independent at the individual level, bias can still result when the grouping matrix $\underset{\sim}{G}$ is not independent of the stochastic disturbance $\underset{\sim}{u}$ (see p. 51-52).

When their description of bias from grouping is translated into more familiar terminology, we find that Feige and Watts actually described the case previously discussed by Blalock (1964) and Hannan (1970; 1971) where the regressand is the basis for group classification. In this case, since $\underset{\sim}{Y}$ is a linear function of $\underset{\sim}{u}$, grouping on $\underset{\sim}{Y}$ ensures that $\underset{\sim}{H}$ and $\underset{\sim}{u}$ are not independent when $\underset{\sim}{Y}$ is the grouping characteristic and thus the estimate from the Y-on-X

regression is biased (see pp. 51-52) for a summary of Blalock's reasoning).

The problem of gauging the magnitude of the divergence remains if the analyst is to systematically choose among alternative grouping methods. The Feige-Watts measure of divergence is based on the difference between $\underset{\sim}{b}$ and $\underset{\sim}{B}$. We summarize the Feige-Watts analysis below, following the Prais-Aitchinson notation and transformation procedures for generating the model at the group level. Relevant equations from our discussion of Prais and Aitchinson are repeated for clarity.

Equation [5.1] with its accompanying assumptions is again the basic model for the ungrouped observations. We have:

$$[5.2] \qquad \underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$$

and

$$[5.3] \qquad \text{cov}(\underset{\sim}{b}) = \sigma_u^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$$

A grouping matrix $\underset{\sim}{G}$ transforms the raw data to a set of m rows; the ith row contains the mean values of the variables for the ith group, I.e., the matrix $[\underset{\sim}{Y},\underset{\sim}{X}]$ is replaced by

$$[\bar{\underset{\sim}{Y}},\bar{\underset{\sim}{X}}] = [\underset{\sim}{G}\underset{\sim}{Y},\underset{\sim}{G}\underset{\sim}{X}] \quad .$$

Recall that

$$\underset{\sim}{H} = \underset{\sim}{G}'(\underset{\sim}{G}\underset{\sim}{G}')^{-1}\underset{\sim}{G} \quad .$$

Hence, the estimates of $\underset{\sim}{\beta}$ and its covariance matrix from grouped data can be written as

$$[5.5] \qquad \underset{\sim}{B} = (\underset{\sim}{X}'\underset{\sim}{H}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{H}\underset{\sim}{Y}$$

and

$$[5.6] \qquad \text{cov}(\underset{\sim}{B}) = \sigma_u^2(\underset{\sim}{X}'\underset{\sim}{H}\underset{\sim}{X})^{-1}$$

The divergence between grouped and ungrouped estimates of $\underset{\sim}{\beta}$,

$$\Delta(\underset{\sim}{H}) = \underset{\sim}{b} - \underset{\sim}{B} \quad ,$$

has a zero mean and variance-covariance matrix equal to

$$\text{cov}[\Delta(\underset{\sim}{H})] = \sigma_u^2[(\underset{\sim}{X}'\underset{\sim}{H}\underset{\sim}{X})^{-1} - (\underset{\sim}{X}'\underset{\sim}{X})^{-1}] \quad .$$

Let $\bar{\underset{\sim}{e}} = \underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{B}$ so that $\bar{\underset{\sim}{e}}'\bar{\underset{\sim}{e}}$ is the sum of squared residuals from the between-groups regression. Assume additionally that the disturbances $\underset{\sim}{u}$ are normally distributed. Then, according to Feige and Watts, the quadratic forms

$$Q_1 = \frac{\Delta(\underset{\sim}{H})[(\underset{\sim}{X}'\underset{\sim}{H}\underset{\sim}{X})^{-1} - (\underset{\sim}{X}'\underset{\sim}{X})^{-1}]\Delta(\underset{\sim}{H})}{\sigma_u^2}$$

and

$$Q_2 = \frac{\bar{\underset{\sim}{e}}'\bar{\underset{\sim}{e}}}{\sigma_u^2}$$

are distributed as χ^2 with k and $m-k$ degrees of freedom, respectively.

Feige and Watts claim that if the model is correctly specified and $\underset{\sim}{H}$ and $\underset{\sim}{u}$ are independent,

[5.8] $$\Gamma = \frac{(Q_1/k)}{[Q_2/(m-k)]}$$

is distributed as F with k and $m-k$ degrees of freedom. Values of Γ beyond the critical values of the F-distribution indicate differences between estimators that cannot be attributed to sampling error. They associate good grouping methods with small Γ values.

The Feige-Watts efficiency criterion is similar to the one that Prais and Aitchinson derived. (See Equation [5.7].) Feige and Watts remove the influence of the constant term, as no information is lost in estimating this parameter. Their efficiency measure is

$$[5.9] \qquad \Gamma = \frac{1}{k-1}\{tr[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{\underset{\sim}{H}}\underset{\sim}{X}] - 1\} \quad ,$$

where $tr[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{\underset{\sim}{H}}\underset{\sim}{X}]$ is the sum of the diagonal elements of the matrix whose entries are the ratios of between-group sums of squares and cross-products to total sums of squares and cross-products. Thus Feige and Watts also recommended forming groups homogeneous with respect to the independent variables in the analysis to minimize loss of efficiency.

To illustrate their findings, Feige and Watts examined twenty regression equations generated from income and dividend invormation provided by 5,393 banks to the Federal Reserve System. The seven grouping rules they used included a random procedure and geographic and financial asset indices. There were also three levels of aggregation -- slight (3 observations per group), moderate (30 observations) and drastic (100 observations). Thus twenty-one grouping methods were possible for each equation although the article only discussed a few.

Certain of the Feige-Watts equations were quite sensitive to the choice of grouping rule and level of aggregation. The reported Γ values ranged from .02 to 84.96. For one equation, all the Γ values were significant at the .05 level, while grouping produced no significant results for other equations. The efficiency indices ranged from .038 to .689, with systematic grouping serving much better than random grouping. In every case, slight aggregation was superior to moderate or drastic aggregation in terms of bias and efficiency. Thus a large number of groups again proved to be desirable.

Otherwise, the Feige-Watts examples demonstrate the tradeoff between efficiency and bias. Random grouping is inefficient but unbiased. Systematic grouping raises the likelihood of misspecification

and grouping bias, but improves efficiency.

It is worth noting that the test that Feige and Watts propose for devergence (Equation 5.8) may not be the most appropriate in this instance. It may well be that the numerator and the denominator are not independent of each other. Furthermore, there is an inherent asymmetry in that the denominator is based solely on the aggregate residuals whereas the numerator is a function of both ungrouped and aggregated information.

The traditional F-test for differences in regression models takes the form:

$$F = \frac{(R_F^2 - R_R^2)/(df_F - df_R)}{(1-R_F^2)/(N-df_F)}$$

where

R_F^2 = squared multiple correlation for the so-called "full" model (the more inclusive model)

R_R^2 = squared multiple correlation for the "restricted" model

and

df_F, df_R = degrees of freedom for the full and restricted models, respectively

There is no recognizable standard for interpreting the comparison of individual-level and aggregate regression models in this fashion. Intuitively, however, it is appealing to associate the individual-level model with the "full" model above and the aggregate with the "restricted". If this interpretation is defensible, then the residual sum of squares from the individual-level regression ($\underline{e}'\underline{e}$, where $\underline{e} = \underline{Y} - \underline{X}\underline{b}$) would seem to be more appropriate than Feige and Watts' choice for the denominator. This is a problem worth exploring further, but it is outside the domain of the present inquiry.

II. The "Structural Equations" Approach in the Two-Predictor Case

The Haitovsky and Feige-Watts conclusions require elaboration since neither presented a way to detect which subset of estimators is biased by grouping. Our analysis of the multiple-regressor case departs from the previous work. First, we specify the order of all variables; the grouping variable is treated as prior to other variables to which it relates. Secondly, each regression coefficient is considered separately. This method, though more cumbersome than a matrix approach, enables us to determine whether the relations of the grouping variable to the regressors and regressand provide clues as to which subset of estimates will exhibit bias. If this strategy works, we will be able to state general principles for determining which estimates are biased for any number of regressors.

We follow a procedure similar to that used in Section IV of Chapter 3 with the bivariate case. A multiple-regression model with two regressors (X and W) is modified by incorporation Z and by specifying the structure among Y , X , W , and Z . This four-variable structural model is then represented by simultaneous equations describing the relations of Y to X , W , and Z , of X to W and Z , and of W to Z .

Formulas for $\beta_{YX \cdot W}$ and $\beta_{YW \cdot X}$ are presented in terms of the parameters of the structural equations at both individual and group levels. The formulas are appropriate for the case when the sample equals the population and under certain conditions for other sampling designs. Any difference between coefficients from grouped and ungrouped data is once more attributed to the effects of grouping.

A. The Regression Equation with Two Regressors

The equation relating Y to X and W is

[5.10] $$Y = \alpha + \beta_{YX \cdot W} X + \beta_{YW \cdot X} W + u \quad ,$$

where

[5.11a] $$\beta_{YX \cdot W} = \frac{\sigma_{YX}\sigma_W^2 - \sigma_{YW}\sigma_{XW}}{\sigma_X^2\sigma_W^2 - (\sigma_{XW})^2}$$

and

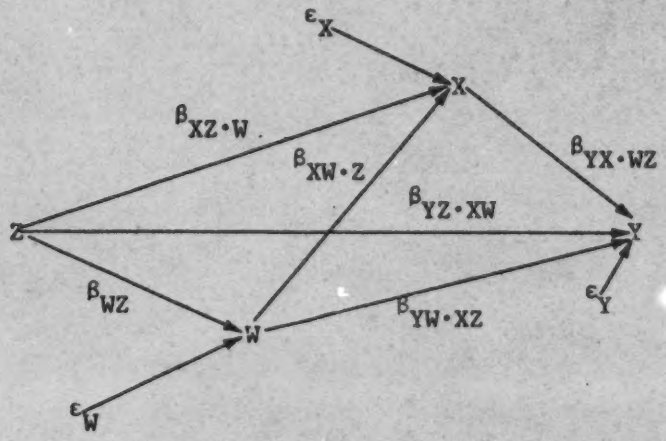
[5.11b] $$\beta_{YW \cdot X} = \frac{\sigma_{YW}\sigma_X^2 - \sigma_{YX}\sigma_{XW}}{\sigma_X^2\sigma_W^2 - (\sigma_{XW})^2}$$

Assumptions B1-B4 still apply so that u is independent of X and W . The object of the investigation is to estimate $\beta_{YX \cdot W}$ and $\beta_{YW \cdot X}$ from equation [5.10] using grouped data.

B. Modified Structure with Z Incorporated

The next step is to constrain the model by specifying a structure among Y , X , W , and Z . As before (see page 53), we treat Z as prior to Y , X , and W . We also assume that W is prior to X and Y .

The path diagram of the structure is



In the diagram, ε_Y is the disturbance term representing all determiners of Y not linearly related to X , W , and Z ; ε_X represents all determiners of X not linearly related to W and Z ; and ε_W represents all

determiners of W not linearly related to Z. $\beta_{YX\cdot W}$, $\beta_{YW\cdot XZ}$, $\beta_{YZ\cdot XW}$, $\beta_{XW\cdot Z}$, $\beta_{XZ\cdot W}$, and β_{WZ} are path regression coefficients.

The structure generates these simultaneous equations:

[5.12a] $Y = \alpha_Y + \beta_{YX\cdot WZ}X + \beta_{YW\cdot XZ}W + \beta_{YZ\cdot XW}Z + \varepsilon_Y$,

[5.12b] $X = \alpha_X + \beta_{XW\cdot Z}W + \beta_{XZ\cdot W}Z + \varepsilon_X$,

[5.12c] $W = \alpha_W + \beta_{WZ}Z + \varepsilon_W$.

Once again, $\beta_{YX\cdot WZ}$, $\beta_{YW\cdot XZ}$, $\beta_{YZ\cdot XW}$, $\beta_{XW\cdot Z}$, $\beta_{XZ\cdot W}$, and β_{WZ} are regression parameters; α_Y, α_X, and α_W are intercepts; and ε_Y, ε_X, and ε_W are disturbance terms. ε_Y is assumed independent of X, W, ε_X, and ε_W. ε_X is assumed independent of W, Z, and ε_W. ε_W is assumed independent of Z. We also assume that the disturbance terms are homoscedastic and independent, as in the single-regressor case.

Besides the intercepts, there are ten parameters: σ_Z^2, $\sigma_{\varepsilon_W}^2$, $\sigma_{\varepsilon_X}^2$, $\sigma_{\varepsilon_Y}^2$, and the six regression coefficients. Rewriting equations [5.12a, b, c] in terms of these parameters, we have

[5.13a] $Y = \alpha_Y + \beta_{YX\cdot WZ}[\alpha_X + \beta_{XW\cdot Z}(\alpha_W + \beta_{WZ}Z + \varepsilon_W) + \beta_{XZ\cdot W}Z + \varepsilon_X]$
$+ \beta_{YW\cdot XZ}(\alpha_W + \beta_{WZ}Z + \varepsilon_W) + \beta_{YZ\cdot XW}Z + \varepsilon_Y$,

[5.13b] $X = \alpha_X + \beta_{XW\cdot Z}(\alpha_W + \beta_{WZ}Z + \varepsilon_W) + \beta_{XZ\cdot W}Z + \varepsilon_X$,

[5.13c] $W = \alpha_W + \beta_{WZ}Z + \varepsilon_W$.

Reduced-form expressions for variances and covariance are

[5.14a] $\sigma_X^2 = (\beta_{XW\cdot Z}^2\beta_{WZ}^2 + \beta_{XZ\cdot W}^2 + 2\beta_{XZ\cdot W}\beta_{XW\cdot Z}\beta_{WZ})\sigma_Z^2$
$+ \beta_{XW\cdot Z}^2\sigma_{\varepsilon_W}^2 + \sigma_{\varepsilon_X}^2$.

[5.14b] $$\sigma_W^2 = \beta_{WZ}^2\sigma_Z^2 + \sigma_{\varepsilon_X}^2 \quad ,$$

[5.14c] $$\sigma_{YX} = (\beta_{YX\cdot WZ}\beta_{XW\cdot Z}^2\beta_{WZ}^2 + \beta_{YX\cdot WZ}\beta_{XZ\cdot W}^2 + 2\beta_{YX\cdot WZ}\beta_{XZ\cdot W}\beta_{XW\cdot Z}\beta_{WZ} + \beta_{YW\cdot XZ}\beta_{XW\cdot Z}\beta_{WZ}^2 + \beta_{YW\cdot XZ}\beta_{XZ\cdot W}\beta_{WZ} + \beta_{YZ\cdot XW}\beta_{XZ\cdot W} + \beta_{YZ\cdot XW}\beta_{XW\cdot Z}\beta_{WZ})\sigma_Z^2 + (\beta_{YX\cdot WZ}\beta_{XW\cdot Z}^2 + \beta_{YW\cdot XZ}\beta_{XW\cdot Z})\sigma_{\varepsilon_W}^2 + \beta_{YX\cdot WZ}\sigma_{\varepsilon_X}^2 \quad ,$$

[5.14d] $$\sigma_{YW} = (\beta_{YX\cdot WZ}\beta_{XZ\cdot W}\beta_{WZ} + \beta_{YX\cdot WZ}\beta_{XW\cdot Z}\beta_{WZ}^2 + \beta_{YW\cdot XZ}\beta_{WZ}^2 + \beta_{YZ\cdot XW}\beta_{WZ})\sigma_Z^2 + (\beta_{YX\cdot WZ}\beta_{XW\cdot Z} + \beta_{YW\cdot XZ})\sigma_{\varepsilon_W}^2 \quad ,$$

[5.14e] $$\sigma_{XW} = (\beta_{XW\cdot Z}\beta_{WZ}^2 + \beta_{XZ\cdot W}\beta_{WZ})\sigma_Z^2 + \beta_{XW\cdot Z}\sigma_{\varepsilon_W}^2 \quad .$$

The reduced-form equations and variance-covariance expressions can be used to derive equations stating $\beta_{YX\cdot W}$ and $\beta_{YW\cdot X}$ in terms of the known parameters. By substitution and rearrangement, we arrive at the desired equations:

[5.15a] $$\beta_{YX\cdot W} = \beta_{YX\cdot WZ} + \beta_{YZ\cdot XW}\beta_{XZ\cdot W}\left[\frac{\sigma_Z^2\sigma_{\varepsilon_W}^2}{\beta_{XZ\cdot W}^2\sigma_Z^2\sigma_{\varepsilon_W}^2 + (\beta_{WZ}^2\sigma_Z^2 + \sigma_{\varepsilon_W}^2)\sigma_{\varepsilon_X}^2}\right]$$

and

[5.15b] $$\beta_{YW\cdot X} = \beta_{YW\cdot XZ} + \beta_{YZ\cdot XW}\sigma_Z^2\left[\frac{\beta_{WZ}\sigma_{\varepsilon_X}^2 - \beta_{XZ\cdot W}\beta_{XW\cdot Z}\sigma_{\varepsilon_W}^2}{\beta_{XZ\cdot W}^2\sigma_Z^2\sigma_{\varepsilon_W}^2 + (\beta_{WZ}^2\sigma_Z^2 + \sigma_{\varepsilon_W}^2)\sigma_{\varepsilon_X}^2}\right]$$

C. Equations Based on Grouped Observations

The equations in Sections II.A and II.B are applicable to the population of ungrouped observations. There is a parallel set of equations for the population of grouped observations.

The initial model for the regression of $\bar{Y}$ on $\bar{X}$ and $\bar{W}$ can be written

[5.16] $\bar{Y} = \alpha + \beta_{\bar{Y}\bar{X}\cdot\bar{W}}\bar{X} + \beta_{\bar{Y}\bar{W}\cdot\bar{X}}\bar{W} + \bar{u}_Y$.

In [5.16], each term is the grouped counterpart of a term for equation [5.10]. $\beta_{\bar{Y}\bar{X}\cdot\bar{W}}$ and $\beta_{\bar{Y}\bar{W}\cdot\bar{X}}$ are the regression coefficients for the grouped observations. Under certain conditions to be discussed below, $\beta_{YX\cdot W} = \beta_{\bar{Y}\bar{X}\cdot\bar{W}}$ and $\beta_{YW\cdot X} = \beta_{\bar{Y}\bar{W}\cdot\bar{X}}$.

The simultaneous equations pertinent to grouped data are given by

[5.17a] $\bar{Y} = \alpha_Y + \beta_{YX\cdot WZ}\bar{X} + \beta_{YW\cdot XZ}\bar{W} + \beta_{YZ\cdot XW}\bar{Z} + \bar{\epsilon}_Y$,

[5.17b] $\bar{X} = \alpha_X + \beta_{XW\cdot Z}\bar{W} + \beta_{XZ\cdot W}\bar{Z} + \bar{\epsilon}_X$,

[5.17c] $\bar{W} = \alpha_W + \beta_{WZ}\bar{Z} + \bar{\epsilon}_W$.

The regression coefficients are given by

[5.18a] $$\beta_{\bar{Y}\bar{X}\cdot\bar{W}} = \beta_{YX\cdot WZ} + \beta_{YZ\cdot XW}\beta_{XZ\cdot W}\left[\frac{\sigma^2_{\bar{Z}}\sigma^2_{\bar{\epsilon}_W}}{\beta^2_{XZ\cdot W}\sigma^2_{\bar{Z}}\sigma^2_{\bar{\epsilon}_W} + (\beta^2_{WZ}\sigma^2_{\bar{Z}} + \sigma^2_{\bar{\epsilon}_W})\sigma^2_{\bar{\epsilon}_X}}\right]$$

and

[5.18b] $$\beta_{\bar{Y}\bar{W}\cdot\bar{X}} = \beta_{YW\cdot XZ} - \beta_{YZ\cdot XW}\sigma^2_{\bar{Z}}\left[\frac{\beta_{WZ}\sigma^2_{\bar{\epsilon}_X} - \beta_{XZ\cdot W}\beta_{XW\cdot Z}\sigma^2_{\bar{\epsilon}_X}}{\beta^2_{XZ\cdot W}\sigma^2_{\bar{Z}}\sigma^2_{\bar{\epsilon}_W} + (\beta^2_{WZ}\sigma^2_{\bar{Z}} + \sigma^2_{\bar{\epsilon}_W})\sigma^2_{\bar{\epsilon}_X}}\right] .$$

Thus the only difference between equations of the grouped and ungrouped regression coefficients ([5.18a, b] compared with [5.15a, b]) is the replacement of population variances by between-group variances.

D. Bias Formulas

Let $b_{YX\cdot W}$ and $b_{YW\cdot X}$ be least-squares estimators of $\beta_{YX\cdot W}$ and $\beta_{YW\cdot X}$, respectively. Also, let $B_{\bar{Y}\bar{X}\cdot\bar{W}}$ and $B_{\bar{Y}\bar{W}\cdot\bar{X}}$ be least-squares estimators of $\beta_{\bar{Y}\bar{X}\cdot\bar{W}}$ and $\beta_{\bar{Y}\bar{W}\cdot\bar{X}}$. Under assumptions B1—B4,

$$E(b_{YX\cdot W}) = \beta_{YX\cdot W}\ ,\ E(b_{YW\cdot X}) = \beta_{YW\cdot X}$$

and

$$E(B_{\bar{Y}\bar{X}\cdot\bar{W}}) = \beta_{\bar{Y}\bar{X}\cdot\bar{W}}\ ,\ E(b_{\bar{Y}\bar{W}\cdot\bar{X}}) = \beta_{\bar{Y}\bar{W}\cdot\bar{X}}\ .$$

That is, all four estimators are unbiased for their own coefficients. But since the investigator is interested in relations at the individual level, his estimates based on grouped data are biased unless $\beta_{\bar{Y}\bar{X}\cdot\bar{W}} = \beta_{YX\cdot W}$ and $\beta_{\bar{Y}\bar{W}\cdot\bar{X}} = \beta_{YW\cdot X}$. We add a subscript to θ to indicate the regressor under consideration; that is, θ_W will denote the bias in estimating $\beta_{YW\cdot X}$ from grouped data and θ_X will denote the bias from estimating $\beta_{YX\cdot W}$. From equations [5.15a, b] and [5.18a, b], we get

$$[5.19a]\quad \theta_X = E(B_{\bar{Y}\bar{X}\cdot\bar{W}}) - E(b_{YX\cdot W})$$

$$= \beta_{YZ\cdot XW}\beta_{XZ\cdot W}\left\{\frac{\sigma^2_{\bar{Z}}\sigma^2_{\bar{\epsilon}_W}(\beta^2_{WZ}\sigma^2_Z+\sigma^2_{\epsilon_W})\sigma^2_{\epsilon_X} - \sigma^2_Z\sigma^2_{\epsilon_W}(\beta^2_{WZ}\sigma^2_{\bar{Z}}+\sigma^2_{\bar{\epsilon}_W})\sigma^2_{\bar{\epsilon}_X}}{[\beta^2_{XZ\cdot W}\sigma^2_Z\sigma^2_{\epsilon_W}+(\beta^2_{WZ}\sigma^2_Z+\sigma^2_{\epsilon_W})\sigma^2_{\epsilon_X}][\beta^2_{XZ\cdot W}\sigma^2_{\bar{Z}}\sigma^2_{\bar{\epsilon}_W} + (\beta^2_{WZ}\sigma^2_{\bar{Z}}+\sigma^2_{\bar{\epsilon}_W})\sigma^2_{\bar{\epsilon}_X}]}\right\}$$

and

$$[5.19b]\quad \theta_W = E(B_{\bar{Y}\bar{W}\cdot\bar{X}}) - E(b_{YW\cdot X})$$

$$= \beta_{YZ\cdot XW} \left\{ \frac{\left[\begin{array}{l} (\beta^2_{XZ\cdot W}\beta_{WZ} + \beta_{XW\cdot Z}\beta_{XZ\cdot W}\beta^2_{WZ})(\sigma^2_{\bar{\varepsilon}_X}\sigma^2_{\bar{\varepsilon}_W} - \sigma^2_{\varepsilon_X}\sigma^2_{\bar{\varepsilon}_W})\sigma^2_{\bar{Z}}\sigma^2_{Z} \\ + \beta_{WZ}\sigma^2_{Z}\sigma^2_{\varepsilon_X}\sigma^2_{\bar{\varepsilon}_X}\sigma^2_{\varepsilon_W} - \beta_{XW\cdot Z}\beta_{XZ\cdot W}\sigma^2_{\bar{Z}}\sigma^2_{\varepsilon_X}\sigma^2_{\varepsilon_W}\sigma^2_{\bar{\varepsilon}_W} \\ - \beta_{WZ}\sigma^2_{Z}\sigma^2_{\varepsilon_X}\sigma^2_{\bar{\varepsilon}_X}\sigma^2_{\bar{\varepsilon}_W} + \beta_{XW\cdot Z}\beta_{XZ\cdot W}\sigma^2_{Z}\sigma^2_{\varepsilon_X}\sigma^2_{\varepsilon_W}\sigma^2_{\bar{\varepsilon}_W} \end{array} \right]}{\left[\begin{array}{l} [\beta^2_{XZ\cdot W}\sigma^2_{Z}\sigma^2_{\varepsilon_W} + (\beta^2_{WZ}\sigma^2_{Z} + \sigma^2_{\varepsilon_W})\sigma^2_{\varepsilon_X}][\beta^2_{XZ\cdot W}\sigma^2_{\bar{Z}}\sigma^2_{\bar{\varepsilon}_W} \\ + (\beta^2_{WZ}\sigma^2_{\bar{Z}} + \sigma^2_{\bar{\varepsilon}_W})\sigma^2_{\bar{\varepsilon}_X}] \end{array} \right]} \right\} .$$

These bias formulas are complicated, especially for the prior regressor W . However, it is clear that there will be no bias so long as the grouping variable has no direct relation to the dependent variable ($\beta_{YZ\cdot XW} = 0$) .

III. The Taxonomy for Two Regressors

A taxonomy can be generated by setting various combination of $\beta_{YZ\cdot XW}$ and β_{WZ} equal to zero. This generates $2 \times 2 \times 2 = 8$ categories of grouping variables:

(1) Z directly related to Y , X , and W ($\beta_{XZ\cdot XW} \neq 0$, $\beta_{XZ\cdot W} \neq 0$, $\beta_{WZ} \neq 0$).

(2) Z directly related to Y and X , but not to W ($\beta_{YZ\cdot XW} \neq 0$, $\beta_{XZ\cdot W} \neq 0$, $\beta_{WZ} = 0$).

(3) Z directly related to Y and W , but not to X ($\beta_{YZ\cdot XW} \neq 0$, $\beta_{XZ\cdot W} = 0$, $\beta_{WZ} \neq 0$).

(4) Z directly related to Y , but not to X or W ($\beta_{YZ\cdot XW} \neq 0$, $\beta_{XZ\cdot W} = 0$, $\beta_{WZ} = 0$).

(5) Z directly related to X and W , but not to Y ($\beta_{YZ \cdot XW} = 0$, $\beta_{XZ \cdot W} \neq 0$, $\beta_{WZ} \neq 0$).

(6) Z directly related to X , but not to Y or W ($\beta_{YZ \cdot XW} = 0$, $\beta_{XZ \cdot W} \neq 0$, $\beta_{WZ} = 0$).

(7) Z directly related to W , but not to Y or X ($\beta_{YZ \cdot XW} = 0$, $\beta_{XZ \cdot W} = 0$, $\beta_{WZ} \neq 0$).

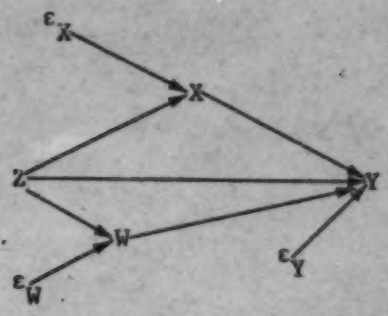
(8) Z not linearly related to Y , X or Z . ($\beta_{YZ \cdot XW}$, $\beta_{XZ \cdot W} = 0$, $\beta_{WZ} = 0$).

As the relation of W to X can also affect bias under certain conditions, it is useful to subdivide each category on the basis of whether $\beta_{XW \cdot Z}$ is non-zero or not. Figure 5.1 presents the sixteen path diagram.

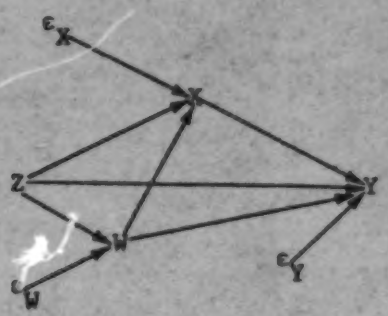
Table 5.2 summarizes the results for bias in the two-regressor case. Grouping by a variable from five of the sixteen subcategories biases the estimate of at least one regression coefficient. There are obvious parallels with the single-regressor case. When there is no direct relation of Z to Y , estimates are unbiased. However, when the grouping variable is directly related to both the dependent variable and a regressor (Categories 2 and 3), the estimate of the coefficient from the regressor is biased when the regressors are correlated. This is analogous to Category I grouping in the bivariate case and the results are the same.

The only result that is not analogous to the bivariate case occurs when $\beta_{XW \cdot Z} \neq 0$ and we estimate the coefficient of the prior regressor. Under this condition, biased estimates $\beta_{YW \cdot X}$ can result when Z is directly related to Y and X even though $\beta_{WZ} = 0$.

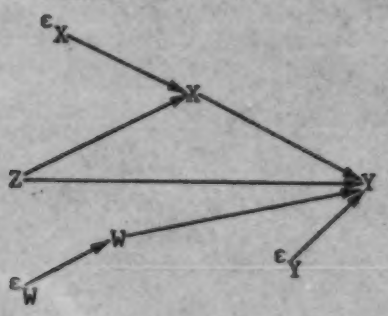
(1) $\beta_{YZ \cdot XW} \neq 0$, $\beta_{XZ \cdot W} \neq 0$, $\beta_{WZ} \neq 0$, $\beta_{XW \cdot Z} = 0$



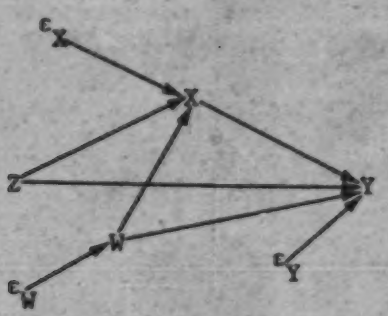
$\beta_{YZ \cdot XW} \neq 0$, $\beta_{XZ \cdot W} \neq 0$, $\beta_{WZ} \neq 0$, $\beta_{XW \cdot Z} \neq 0$



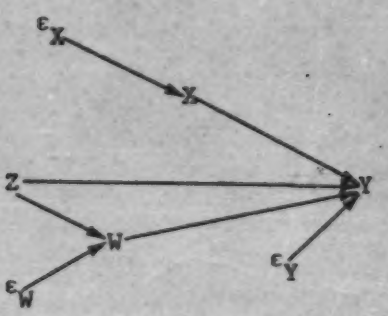
(2) $\beta_{YZ \cdot XW} \neq 0$, $\beta_{XZ \cdot W} \neq 0$, $\beta_{WZ} = 0$, $\beta_{XW \cdot Z} = 0$



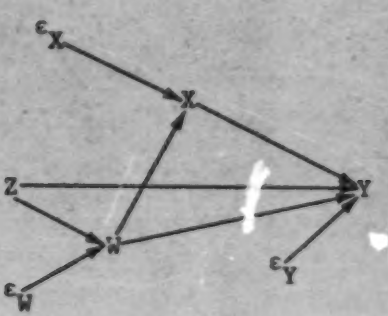
$\beta_{YZ \cdot XW} \neq 0$, $\beta_{XZ \cdot W} \neq 0$, $\beta_{WZ} = 0$, $\beta_{XW \cdot Z} \neq 0$



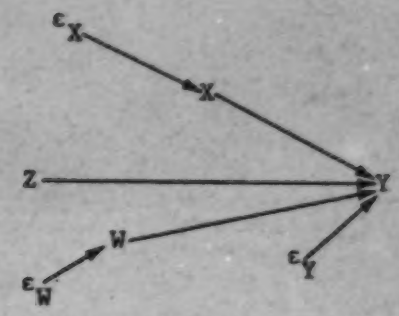
(3) $\beta_{YZ \cdot XW} \neq 0$, $\beta_{XZ \cdot W} = 0$, $\beta_{WZ} \neq 0$, $\beta_{XW \cdot Z} = 0$



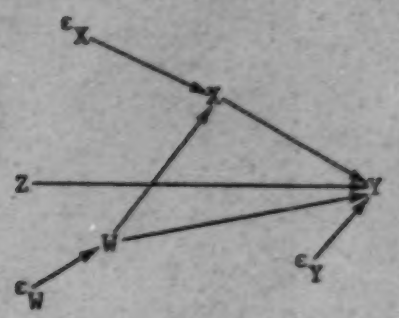
$\beta_{YZ \cdot XW} \neq 0$, $\beta_{XZ \cdot W} = 0$, $\beta_{WZ} \neq 0$, $\beta_{XW \cdot Z} \neq 0$



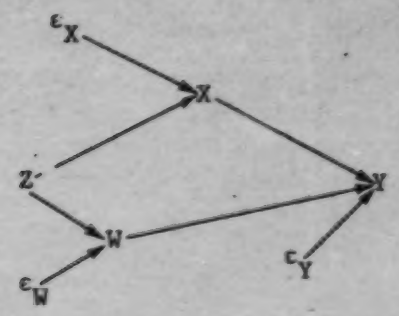
(4) $\beta_{YZ \cdot XW} \neq 0$, $\beta_{XZ \cdot W} = 0$, $\beta_{WZ} = 0$, $\beta_{XW \cdot Z} = 0$



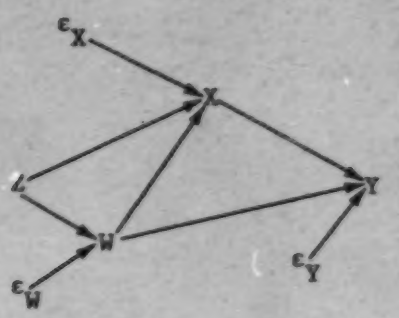
$\beta_{YZ \cdot XW} \neq 0$, $\beta_{XZ \cdot W} = 0$, $\beta_{WZ} = 0$, $\beta_{XW \cdot Z} \neq 0$



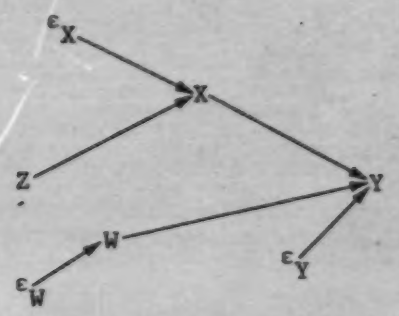
(5) $\beta_{YZ \cdot XW} = 0$, $\beta_{XZ \cdot W} \neq 0$, $\beta_{WZ} \neq 0$, $\beta_{XW \cdot Z} = 0$



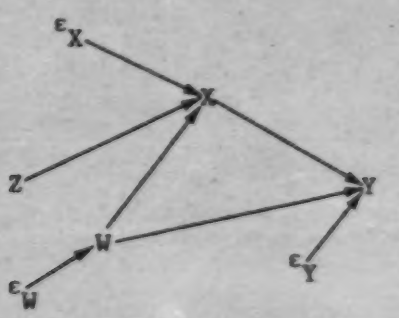
$\beta_{YZ \cdot XW} = 0$, $\beta_{XZ \cdot W} \neq 0$, $\beta_{WZ} \neq 0$, $\beta_{XW \cdot Z} \neq 0$



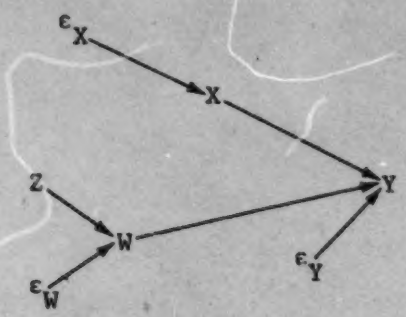
(6) $\beta_{YZ \cdot XW} = 0$, $\beta_{XZ \cdot W} \neq 0$, $\beta_{WZ} = 0$, $\beta_{XW \cdot Z} = 0$



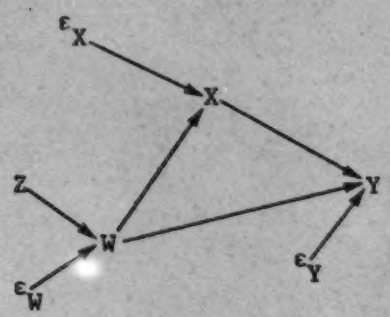
$\beta_{YZ \cdot XW} = 0$, $\beta_{XZ \cdot W} \neq 0$, $\beta_{WZ} = 0$, $\beta_{XW \cdot Z} \neq 0$



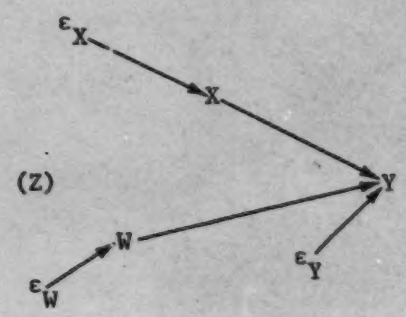
(7) $\beta_{YZ \cdot XW} = 0$, $\beta_{XZ \cdot W} = 0$, $\beta_{WZ} \neq 0$, $\beta_{XW \cdot Z} = 0$



$\beta_{YZ \cdot XW} = 0$, $\beta_{XZ \cdot W} = 0$, $\beta_{WZ} \neq 0$, $\beta_{XW \cdot Z} \neq 0$



(8) $\beta_{YZ \cdot XW} = 0$, $\beta_{XZ \cdot W} = 0$, $\beta_{WZ} = 0$, $\beta_{XW \cdot Z} = 0$



$\beta_{YZ \cdot XW} = 0$, $\beta_{XZ \cdot W} = 0$, $\beta_{WZ} = 0$, $\beta_{XW \cdot Z} \neq 0$

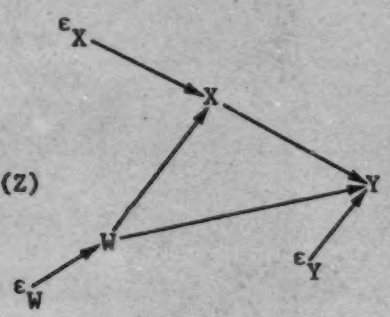


Figure 5.1. Path diagrams for the subcategories of the taxonomy in the two-regressor case.

Table 5.2. Presence of bias from grouping as a function of taxonomic subcategory in the two-regressor case.

Category	Values of Parameters				Bias in Coefficients[a]	
	$\beta_{YZ \cdot XW}$	$\beta_{XZ \cdot W}$	β_{WZ}	$\beta_{XW \cdot Z}$	$\beta_{YX \cdot W}$	$\beta_{YW \cdot X}$
1	$\neq 0$	$\neq 0$	$\neq 0$	0	*	*
1	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	*	*
2	$\neq 0$	$\neq 0$	0	0	*	
2	$\neq 0$	$\neq 0$	0	$\neq 0$	*	*
3	$\neq 0$	0	$\neq 0$	0		*
3	$\neq 0$	0	$\neq 0$	$\neq 0$		*
4	$\neq 0$	0	0	0		
4	$\neq 0$	0	0	$\neq 0$		
5	0	$\neq 0$	$\neq 0$	0		
5	0	$\neq 0$	$\neq 0$	$\neq 0$		
6	0	$\neq 0$	0	0		
6	0	$\neq 0$	0	$\neq 0$		
7	0	0	$\neq 0$	0		
7	0	0	$\neq 0$	$\neq 0$		
8	0	0	0	0		
8	0	0	0	$\neq 0$		

[a] * = Estimator of regression coefficient from grouped data is biased.

IV. Implications of Findings

Our taxonomic approach clarifies certain questions raised by earlier investigations of multiple regression. We have shown that bias can result for only a subset of regression coefficients. In fact, the conditions under which the estimator of a particular coefficient will be biased can now be specified.

Much has been left unsaid about the practical consequences of grouping in the multiple-regressor case. Bias in estimating at least some coefficients is highly likely unless groups are formed randomly. With non-random grouping, the investigator may group on a variable which is prior to all others. Otherwise, he introduces bias in estimating some coefficients by grouping jointly on the dependent variable and posterior regressors.

The "structural equations" approach does enable the investigator to determine which estimators are biased, but the procedures quickly become cumbersome as more independent variables are included. More work is needed to determine the utility of this approach, especially when compared with the procedures developed by Feige and Watts.

CHAPTER 6

EMPIRICAL EXAMPLES IN THE SINGLE-REGRESSOR CASE

So far we have considered ways of predicting how various grouping procedures affect the estimation of simple linear regression coefficients. It seems appropriate at this point to demonstrate how well our predictions conform to empirical results. Information collected on incoming freshmen at a large Midwestern university serves as the data base for this investigation. Of 300 measures of abilities, attitudes, and interests collected originally, approximately 20 will be used. Persons with missing information on any of these variables are dropped from consideration.

First we describe the relevant variables and the form in which they enter the analysis. Next a simple linear regression model is hypothesized, and the regression slope and its standard error are estimated from the ungrouped observations.

The data are then grouped. We vary the relation of the grouping variable to the dependent and independent variables, the number of groups formed, and the distribution of observations among the groups. Estimates of the regression slope and its standard error are then calculated from the grouped observations for each grouping variable. The difference between the observed grouped and ungrouped slope estimates (bias) is then compared with that predicted from the formulas of Chapter 3. Indices of efficiency are also presented. We then discuss the potential utility of composite estimates, formed from the estimates generated by different grouping characteristics, in making inferences about the individual-level relations.

I. Description of Data

All incoming freshmen at a large Midwestern university were administered an achievement battery consisting of arithmetic, mathematics and reading comprehension subtests during their orientation session prior to entering the university. On the last day of orientation, each student was asked to complete inventories assessing his personal history, his interests, his expectations regarding his university experience, and his opinions about selected social and academic issues. In our example, this information was later combined with data from admissions applications and with scores from the Scholastic Aptitude Test (SAT).

A. Identification of Variables

We focus on the relation of achievement (X) to self-appraisal (Y) of academic abilities and of SAT(X) to achievement (Y). Each student's total score on the achievement test battery (ACH) represents his achievement level. The indicator (SRAA) of self-rated academic abilities is a weighted composite of responses to ten questions (Table 6.1) asking the student to rate his abilities of his work in different academic areas.

The weights of variables entering the composite self-rating were determined by variable loadings on the first factor from a principal components analysis. The weights were relatively uniform except that mathematics ability and scientific ability had small weights. Thus the analysis leads us to equate students' preceptions of their academic ability mainly with their verbal communications skills.

Subgroups of students were formed on the basis of their SAT, ACH and SRAA scores. Students were classified into subgroups according to the highest two digits of their ACH scores (ACH2, 10 groups: 31-39, 40-49, ..., 110-119; 120), of their SAT scores (SAT2, 13 groups: 400-499, 500-599, ..., 1500-1599; 1600), and of their SRAA scores (SRAA2, 5 groups:

Table 6.1. Questions included in composite self-appraisal of academic abilities (SRAA).

Use the instructions below for answering questions 1 through 4:

"Rate yourself on each of the following traits as you really think you are when compared with the average student of your own age."

Scale: A. Lowest 10%
B. Below average
C. Average
D. Above average
E. Highest 10%

1. Academic ability
2. Mathematical ability
3. Self confidence (intellectual)
4. Writing ability

Use the instructions below for answering questions 5 through 8:

"Rate yourself on how competent you feel you are when compared to other freshmen at the university."

Scale: A. Lowest 10%
B. Below average
C. Average
D. Above average
E. Highest 10%

5. Overall scholarship
6. Scientific ability
7. Reading skills
8. Intellectual self-confidence
9. Where do you think you are likely to rank with respect to grades in your freshman class while in college?

Scale: A. Among the highest 10%
B. Above average
C. About average
D. Below average
E. Among the lowest 10%

10. Forget for a moment how others grade your work. In general, what is your own opinion as to how good your academic work will be?

Scale: A. Excellent
B. Very good
C. About Average
D. Somewhat below average
E. Much below average

-2.99 to -2.00, -1.99 to -1.00, -0.99 to 0.99, 1.00 to 1.99; 2.00 to 2.99). ACH2, SAT2, and SRAA2, then, were the grouping variables based on ACH, SAT, and SRAA, respectively.

The remaining grouping variables were selected according to the following criteria:

1. The variable has appeared frequently in studies of the relations among academic self-appraisal, achievement, and aptitude (e.g., parental income, parental education, parents' educational aspirations for their children).

2. Alternatively, the frequency distribution of the variable and its pattern of zero-order correlations with ACH, SAT, and SRAA suggested that it would be a suitable representative of a particular taxonomic category (e.g, number of semesters of high-school physical sciences, student opinion about whether college is worth the effort, and the last 2 digits of the student's identification number).

Table 6.2 lists the grouping variables, ordered according to number of groups formed (except for the two "identification" variables at the top which serve as random numbers in our example).

B. Distributional and Relational Properties of the Variables

Table 6.3 lists for each study variable the mean, standard deviation, and skewness coefficient and zero-order correlations with SRAA, ACH, and SAT. Only the 2,676 students with complete information on all variables are used here and later.[1]

---

[1]After the bulk of the analyses was completed, it was discovered that there were missing observations on the grouping characteristics CLIMP, COLEFF, and QCJOB. In addition certain modifications were made in the response categories of ANTDEG. In its original form, ANTDEC formed nine groups. In the results reported here, however, students responding "Other (9)" were dropped, and students anticipating any professional degree beyond the masters level (responses 5, 6, 7, and 8) were collapsed into a single group numbered "5". The sizes of the subsamples defined by the acceptable responses to CLIMP, COLEFF, QCJOB, and the modified ANTDEG were 2,632, 2,669, 2,637, and 2,646,

Table 6.2. Information on grouping variables.

Variable Identification	Description	Number of Groups After Aggregation
ID2	Last 2 digits of student identification	100
ID1	Last digit of student identification	10
HSGPA2	High school's report of student's grade point average on a 4-point scale (highest 2 digits)	23
SAT2	Highest 2 digits of Total score from the Scholastic Aptitude Test	13
ACH2	Highest 2 digits of Total score from the Achievement Battery	10
PARINC	Student's best estimate of 1970 parental income before taxes	10
REPGPA	Student's report of average grade in secondary school	7
POPED	Student's report of highest level of formal education obtained by his father	6
ANTDEG	Student's anticipated highest academic degree	5
HSMATH	Student's report of number of semesters of high school mathematics	5
HSPHYS	Student's report of number of semesters of high school physical sciences	5
NOBOOK	Student's report of number of books in the home	5
PARASP	"What is the highest level of education that your parents hope you will complete?"	5
SRAA2	Highest digit and sign of composite academic self-opinion	5
CLIMP	"My grades are markedly better in courses that I see I will need later."	4
COLEFF	"I often wonder if four years of college will really be worth the effort."	4
OCJOB	"I often wish that I were offered a good job now so I wouldn't have to spend four years in college.	4

Table 6.3. Means, standard deviations, and skewness coefficients of study variables, and the zero-order correlations of each variable with SRAA, ACH, and SAT.

Variable Name	Mean	Standard Deviation	Skewness	Correlation with SRAA	ACH	SAT
SRAA	0.008	1.006	.223	1.000	.529	.574
ACH	84.766	15.463	-.364	.529	1.000	.839
SAT	1068.846	177.209	.068	.574	.839	1.000
ID2	49.561	29.126	.003	.019	.020	.008
ID1	4.453	2.865	.011	-.033	-.042	-.047
HSGPA2	3.157	.469	-.067	.370	.535	.488
SAT2	10.235	1.798	.064	.566	.827	.987
ACH2	8.024	1.572	-.333	.522	.983	.827
PARINC	6.308	2.289	-.234	.064	.070	.076
REPGPA	3.203	1.284	.232	-.455	-.490	-.469
POPED	3.987	1.418	-.321	.145	.139	.157
ANTDEG	3.867	.959	.687	.264	.156	.140
HSMATH	4.332	.879	-.260	.202	.479	.346
HSPHYS	2.623	.977	.319	.209	.318	.257
NOBOOK	4.104	.978	-.759	.196	.146	.203
PARASP	4.458	.626	-1.523	.172	.066	.087
SRAA2	.005	.689	.399	.885	.476	.520
CLIMP	2.201	.821	.304	.074	.147	.165
COLEFF	2.695	.951	-.209	.189	.134	.114
QCJOB	3.330	.821	-1.151	.199	.105	.118

The variables based on the student's identification number (ID2 and ID1) have rectangular distributions as expected. Their intercorrelations with the main variables are close to zero. They satisfactorily represent Category IV ("random") grouping.

In this sample, parental income (PARINC) is weakly related to achievement, aptitude, and academic self-ratings, with correlations not much larger than those from the essentially random ID variables (POPED and NOBOOK).

Anticipated highest degree (ANTDEG) and parental aspirations (PARASP) correlate moderately with each other (.39), but do not correlate with other grouping variables. Both correlate higher with SRAA than with ACH and SAT, perhaps because of similar biases or sets in all student self-report measures.

The grouping variables generated from ACH, SRAA, and SAT (ACH2, SRAA2, and SAT2) and the indicators of high school grades (HSGPA2 and REPGPA) have substantial correlations with the main variables (ACH, SRAA, and SAT). In general these correlations follow predictable patterns. ACH2 correlates highest with ACH, next highest with SAT. SAT2 correlates highest with SAT, next highest with ACH. SRAA2 correlates highest with SRAA, and the order of its correlations with SAT and ACH is the same as for SRAA. HSGPA2 has stronger correlations with the two total test scores than with academic self-rating. The profile of correlations for REPGPA is flatter than that from HSGPA2, but it maintains the same order of magnitude.

---

[1] respectively. An examination of the means, standard deviations, and intercorrelations of SRAA, ACH, and SAT for these subsamples did not indicate any consistent and important deviations from the estimates based on the entire 2,676 observations.

The predominance of four- and five-choice variables has the advantage of easy convertability to group classifications and the disadvantage of low reliability. The substantive importance of these short scales lies in the diversity of their patterns of correlation with achievement, aptitude, and academic self-rating. As will be shown subsequently, reasonably precise[2] estimates of the relations at the individual level can be obtained by grouping on some of these variables, while grouping by others yields wildly misleading estimates. Determining which characteristics coincide with high precision in empirical data is particularly important at this point in the study of grouping effects.

C. Review of Factors Affecting Within-Category Precision

The mechanisms controlling the comparative precision of estimates from different grouping characteristics within a given category vary according to category. The four key "forces" determining precision within a taxonomic category are (1) the relative strengths of the relations of the grouping variable to the dependent and independent variables, (2) the coarseness of the grouping, (3) the between-groups variation in the independent variable for a given grouping characteristic, and (4) the distribution of the individual observations among the groups. We review briefly the manner in which these forces operate, according to the theory developed earlier.

1. Strengths of Relations of Z to X and Y

The standardized regression coefficients best indicate the strength of relations within a given sample. An "*" is introduced

---

[2]There is no exact formula for the "precision" of estimation. Precise estimates generally combine small bias (in our case, $B_{\overline{YX}} - b_{YX}$) with small mean-squared error [MSE = $(\text{bias})^2 + SE(B_{\overline{YX}})^2$] . Whether bias or mean-squared error is more important in defining precision depends on the purpose for which the estimate will be used.

as a superscript for regression coefficients to denote that the coefficients are standardized in this section. In Category III, according to our theory, variables with the weakest direct relation to Y (small $\beta^*_{YZ\cdot X}$) and the strongest relation to X (large β^*_{XZ}) yield the most precise estimates.

The influence is more complicated in Category I. In general, large β^*_{XZ} and small $\beta^*_{YZ\cdot X}$ lead to greater precision. More can be said if we fix one parameter and vary the other, or consider the ratio of $\beta^*_{YZ\cdot X}$ to β^*_{XZ} :

(a) For fixed β^*_{XZ} of any size, the smaller the value of $\beta^*_{YZ\cdot X}$, the smaller the bias.

(b) For small (less than .2 but significantly different from zero) values of $\beta^*_{YZ\cdot X}$, larger values of β^*_{XZ} lead to smaller bias.

(c) Whenever $\beta^*_{YZ\cdot X}$ exceeds β^*_{XZ} for a Category I variable, a particularly poor estimate of β_{YX} results from grouped observations.

2. Coarseness

The coarseness of grouping, by which we mean the number of groups formed (m) from a fixed number of observations (N), largely determines the efficiency with a Category IV grouping characteristic. The strength of relation of Category IV variables to the main variables is inconsequential; hence, they group observations in an essentially random fashion, and the precision of their estimates is influenced only by m.

Coarseness influences bias and efficiency in other categories to a lesser degree. If two variables Z_1 and Z_2 have similar relations to X and Y ($\beta^*_{XZ_1} = \beta^*_{XZ_2}$, $\beta^*_{YZ_1\cdot X} = \beta^*_{YZ_2\cdot X}$), the one with more

groups is likely to yield estimates with smaller bias and higher efficiency.

3. Between-Groups Variation in X

Large between-groups variance in the independent variable implies small bias and high efficiency. With fixed values of m and relatively constant values of $\beta^{*}_{YZ\cdot X}$ and β^{*}_{XZ}, the grouping variable which maximizes the between-groups variance of X yields the most precise estimate.

4. Distribution of Individual Observations Among the Groups

The mean of the dependent or independent variables in a group with few observations is unstable. Such means can have a disproportionate impact on the estimates from grouped observations. Unpredictable variation of a few group means when m is small is potentially more damaging than the same variation among a large number of observations at the individual level. At the group level, the only observations are the means. Instability in any cell mean has a greater impact on the precision of the parameter estimates than does instability at the individual level. When the observations are not evenly distributed among the groups, precision can be affected.

The four forces do not act independently. It makes little sense to consider the impact of $\sigma_{\bar{X}}$ and ignore the size of β^{*}_{XZ}, or to concentrate on coarseness without considering variation in group size. Thus the investigator must keep in mind that the forces can interact. In the discussion of the empirical data, we will only reluctantly attribute a loss of precision to a single source.

II. Regression of Academic Self-Appraisal on Achievement

As our first example, we regress academic self-appraisal (SRAA = Y)

on achievement (ACH = X). Alternative models of the relation between ACH and SRAA are certainly reasonable. However, we only wish to illustrate the effects of grouping, and the chosen ordering is informative.

At the outset we standardize all variables. The procedure for generating group estimates and judging their precision are invariant with regard to linear transformations of the variables. Once the observations are standardized, the regression coefficient at the individual level (the standardized regression coefficient) is an unbiased estimator of the correlation coefficient; i.e., $E(b^*_{YX}) = \beta^*_{YX} = \rho_{YX}$ in the single regressor case. Thus we obtain estimates of ρ_{YX} when we regress $\bar{Y}$ on $\bar{X}$. Under these circumstances, comparisons of $B^*_{\bar{Y}\bar{X}}$ with b^*_{YX} are checks on the bias in estimating the individual-level correlation coefficient from grouped data. (At this point, we will drop the "*" denoting standardized coefficients since all coefficients in the remainder of the chapter will be generated from data that were initially standardized.).

A. Regression Coefficients from Ungrouped Data

According to the analysis of the 2,676 observations, the equation relating to SRAA(Y) to ACH(X) is

$$SRAA = .529(ACH) .$$

That is, the slope of the regression is $b_{YX} = .529$. (The intercept is essentially 0 since all variables were standardized.). Also,

$$SE(b_{YX}) = .0032$$

and

$R^2_{Y \cdot X} = .281$ (the squared multiple correlation coefficient).

In a study such as this, the investigator usually generalizes beyond the 2,676 students included in the analysis. After all, these students are not even the entire freshmen class entering this university during

the 1971-72 academic year.[3] Apparently, our deletion of subjects did leave a representative sample of the freshmen class.[4]

B. Categorization of Grouping Variables

To classify grouping variables (Z) into taxonomic categories requires information beyond that in Table 6.3. Table 6.4 contains for each Z, estimates of the regression coefficients ($\beta_{YX\cdot Z}$, $\beta_{YZ\cdot X}$, β_{YZ}, β_{XZ}) and their standard errors (in parentheses below). An estimate of the between-groups standard deviation, $\sigma_{\bar{X}}$, of ACH for each of the grouping variables is also given.

The taxonomy introduced in Chapter 3 categorizes on the basis of the magnitude of $\beta_{YZ\cdot X}$ and β_{XZ}. Operationally, for initial categorization, we require that $\hat{\beta}_{YZ\cdot X}$ and $\hat{\beta}_{XZ}$ exceed 3 times their standard errors to be considered significantly different from zero. This rather stringent criterion leads to the following category assignments [Variables within categories are ordered by the number of groups they form (m).]:

---

[3] A total of 5,230 students completed the questionnaires during orientation of the 1971-72 academic year. Other students enrolled without attending orientation or participating in the orientation tests and survey. Students who did not begin Fall term were also excluded from the 5,230 total.

[4] An early computer run (carried out before SRAA was created and before the subtests composing the achievement battery were combined to obtain the total achievement score) based on the 4,241 freshmen with reported SAT scores indicated that our students are like their fellow classmates. The average student in our sample performed slightly better on the SAT (1069 to 1054), about the same on the achievement battery (85 to 84), and had the same high school grade average, and reported a slightly higher parental income. The relationship between SAT and PARINC was somewhat stronger (0.109 compared to 0.076) for the 3,647 students with SAT scores who also reported their parents' income than for the students in our sample. Differences in means, standard deviations, and intercorrelations on other characteristics were minor al

Table 6.4. Estimates of parameters relating ACH(X) and SRAA(Y) to possible grouping variables (Z)[a].

Variable Name	Group Size (m)	Parameter Estimates $\hat{\beta}_{YX\cdot Z}$	$\hat{\beta}_{YZ\cdot X}$	$\hat{\beta}_{XZ}$	$\hat{\beta}_{YZ}$	$\sigma_{\bar{X}}$
ID2	100	.529 (.0164)[b]	.008 (.0164)	.020 (.0193)	.019 (.0193)	.189
ID1	10	.528 (.0164)	-.011 (.0164)	-.042 (.0193)	-.033 (.0193)	.078
HSGPA2	23	.463 (.0193)	.123 (.0193)	.535 (.0163)	.370 (.0180)	.552
SAT2	13	.194 (.0282)	.406 (.0282)	.827 (.0109)	.566 (.0160)	.831
ACH2	10	.460 (.0896)	.070 (.0896)	.983 (.0035)	.522 (.0165)	.984
PARINC	10	.527 (.0164)	.028 (.0164)	.070 (.0193)	.064 (.0193)	.122
REPGPA	7	.403 (.0182)	-.258 (.0182)	-.490 (.0169)	-.455 (.0172)	.510
POPED	6	.519 (.0165)	.073 (.0165)	.139 (.0192)	.145 (.0191)	.150
ANTDEG	5	.499 (.0162)	.186 (.0162)	.156 (.0191)	.264 (.0186)	.159
HSMATH	5	.561 (.0187)	-.066 (.0187)	.479 (.0170)	.202 (.0189)	.489
HSPHYS	5	.515 (.0173)	.046 (.0173)	.318 (.0183)	.209 (.0189)	.365
NOBOOK	5	.511 (.0164)	.122 (.0164)	.146 (.0191)	.196 (.0190)	.148
PARASP	5	.520 (.0162)	.138 (.0162)	.066 (.0193)	.172 (.0190)	.077
SRAA2	5	.139 (.0099)	.819 (.0099)	.476 (.0170)	.885 (.0090)	.481
CLIMP	4	.530 (.0166)	-.003 (.0166)	.147 (.0191)	.074 (.0193)	.163
COLEFF	4	.513 (.0164)	.121 (.0164)	.134 (.0192)	.189 (.0190)	.144
QCJOB	4	.514 (.0163)	.145 (.0163)	.105 (.0192)	.199 (.0190)	.113

[a] All variables have been standardized prior to grouping so that $\sigma_Y = \sigma_X = \sigma_Z = 1$, $\beta_{XZ} = \rho_{XZ}$, and $\beta_{YZ} = \rho_{YZ}$.

[b] Numbers in parentheses are standard errors of the regression coefficients.

	$\lvert\hat{\beta}_{YZ\cdot X}\rvert \geq 3SE(\hat{\beta}_{YZ\cdot X})$		$\lvert\hat{\beta}_{YZ\cdot X}\rvert < 3SE(\hat{\beta}_{YZ\cdot X})$
	Category I		Category III
$\lvert\hat{\beta}_{XZ}\rvert \geq 3SE(\hat{\beta}_{XZ})$	HSGPA2	HSMATH	ACH2
	SAT2	NOBOOK	PARINC
	ANTDEG	PARASP	HSPHYS
	REPGPA	COLEFF	CLIMP
	POPED	QCJOB	
	SRAA2		
	Category II		Category IV
$\lvert\beta_{XZ}\rvert < 3SE(\hat{\beta}_{XZ})$	(NONE)		ID2
			ID1

As we mentioned previously, no characteristics belong to Category II, and the number falling in Category I is large. SRAA2 and ACH2 are special cases within Categories I and III. These, respectively, are the best approximations to what Blalock (1964) and Hannan (1970, 1971; 1972) have called "grouping on the dependent variable" and "grouping on the independent variable".

C. Prediction of Bias from Grouping

A modification of the bias formulas ([3.19'], [3.28'], [3.29], [3.31]) from pages 63 and 64 can be used to predict the bias from grouping for our empirical examples. Remembering that $\sigma_X = \sigma_Z = \sigma_{\bar{Z}} = 1$, our equation for estimating the bias from grouping on a particular Z is given by

[5.1] $$\hat{\theta} = \hat{\beta}_{YZ\cdot X}\hat{\beta}_{XZ}\left(\frac{1 - \hat{\sigma}^2_{\bar{X}}}{\sigma^2_{\bar{X}}}\right).$$

This approximation is particularly good when the sample either equals the population or is very large. The small sample properties of θ are less predictable when both $\beta_{YZ\cdot X}$ and β_{XZ} are non-zero. We have included academic self-appraisal in our example because this type of data is often collected anonymously. If so, we cannot correlate ACH

and SRAA at the individual level. The data collected in this study were completely identified, and thus the results under constraints of anonymity can be compared with the results from completely identified data.

As pointed out in Chapter 1 [discussion of Problem (D)], one way to handle anonymous data is to analyze relations at the group level. For example, students can be asked to indicate their number of semesters of high school mathematics (HSMATH) when they complete the attitude questionnaries anonymously. SRAA and ACH scores can then be grouped according to students' HSMATH responses, and the regression of SRAA on ACH can be estimated from the weighted group means of SRAA and ACH .

To be sure, we are still not able to estimate $B_{YZ\cdot X}$ directly since σ_{XY} cannot be determined. Thus we cannot estimate grouping bias θ . But the estimate of β_{YZ} can be used in place of the unobtainable estimate of $\beta_{YZ\cdot X}$ in the equation for bias. This substitution yields a function of the estimated grouping bias.

[5.2]
$$\hat{\pi} = \left(\frac{\hat{\beta}_{YZ}}{\hat{\beta}_{YZ\cdot X}}\right)\hat{\theta}$$

$$= \hat{\beta}_{YZ}\hat{\beta}_{XZ}\left(\frac{1 - \sigma^2_{\bar{X}}}{\hat{\sigma}^2_{\bar{X}}}\right) .$$

In most cases, enough is known about the covariance of X and Y to determine at least its sign. When β_{YZ} is positive (negative) and β_{XZ} and σ_{XY} have the same sign, β_{YZ} provides an upper (lower) bound for $\beta_{YZ\cdot X}$. When β_{XZ} and σ_{XY} have opposite signs, β_{YZ} becomes a lower (upper) bound. Thus, we expect small $\hat{\pi}$ values to occur with

good estimators of β_{YX} and large $\hat{\pi}$ values with poor estimators.

Table 6.5 lists for each grouping variable, the predicted biases (both $\hat{\theta}$ and $\hat{\pi}$) in estimating the coefficient from the regression of SRAA on ACH. Later, we shall compare these values with the observed biases resulting from grouping.

D. Estimates of Regressions from Different Grouping Methods

Two standards are applied for judging the precision of estimating β_{YX} from data grouped on a given Z. First, estimates of bias and efficiency from grouping on different variables are compared, both within and between categories. These comparisons focus on the effects of within-variable factors on precision and on the relative precision of different categories of variables.

We also examine precision on an absolute scale; i.e., independently of the scaling of ACH and SRAA. To do this, we (a) compare observed and predicted bias from grouping with twice the standard error of its estimate, $SE(B_{\overline{YX}})$; and (b) examine indices of efficiency generated from the ratio of the mean-squared error from ungrouped data to the mean-squared error from a particular grouping. Since these standard indices of efficiency tend to be small due to the coarseness of grouping, we also compare the efficiency of a particular grouping with the efficiency of forming an equal number of groups randomly (m-1/N-1).

1. Relative Precision by Category

Table 6.5 contains estimates of the regression coefficients, their standard errors, the observed and predicted grouping bias, and estimates of the square root of the mean-squared error of each grouping variable. The grouping variables are ordered within categories by the size of the observed bias except for ACH2 and SRAA2,

Table 6.5. Estimates from grouped data of coefficients describing the regression of SRAA on ACH.

Grouping Variable	Number of Groups (m)	$B_{\bar{Y}\bar{X}}$[a]	Bias Observed	Bias Predicted from $\hat{\theta}$	Bias Predicted from $\hat{\pi}$	$SE(B_{\bar{Y}\bar{X}})$[a]	$\sqrt{\widehat{MSE}(B_{\bar{Y}\bar{X}})}$[b]
Category IV							
ID2	100	.558	.029	.004	.040	.0739	.0794
ID1	10	.442	-.087	.075	.225	.1831	.2027
Category III[c]							
ACH2	10	.531	.002	.002	.142	.0615	.0615
PARINC	10	.558	.029	.130	.295	.1314	.1345
HSPHYS	5	.571	.042	.095	.433	.0915	.1294
CLIMP	5	.717	.188	.016	.401	.3971	.4382
Category I[c]							
SRAA2	5	1.853	1.324	1.295	1.507	.0631	1.3255
HSMATH	5	.414	-.115	-.100	.307	.0248	.1176
SAT2	13	.671	.142	.150	.210	.0670	.1570
HSGPA2	23	.702	.173	.150	.451	.0287	.1753
POPED	6	.911	.382	.440	.874	.1626	.4152
REPGPA	7	.917	.388	.360	.635	.0617	.3929
NOBOOK	5	1.334	.805	.800	1.285	.1133	.8129
COLEFF	4	1.461	.932	.765	1.194	.1160	.9392
ANTDEG	5	1.631	1.102	1.117	1.586	.2680	1.1341
QCJOB	4	1.853	1.324	1.188	1.630	.3533	1.3703
PARASP	5	1.946	1.417	1.519	2.048	.7339	1.5958

[a] Estimates from ungrouped data: $b_{YX} = .529$; $SE(b_{YX}) = .0032$.

[b] $\sqrt{\widehat{MSE}(B_{\bar{Y}\bar{X}})} = \sqrt{(\text{OBSERVED BIAS})^2 + [SE(B_{\bar{Y}\bar{X}})]^2}$

[c] With the exception of ACH2 and SRAA2, variables within categories are ordered on the basis of observed bias.

which are listed first in their respective categories.

In general, the estimates conform to our expectations though the bias and mean-squared error (MSE)[5] are enormous for some Category I variables. Category IV grouping yielded estimates with small bias. In fact, only grouping on ACH2 (grouping on the independent variable) gives better precision (small bias and small mean-squared error) than the estimate from ID2 . But it took ten times as many groups to achieve this level of accuracy.

The bias from grouping by ID1 , the other Category IV variable, is three times as large as the bias from grouping on ID2 . Its estimated MSE is more than six times larger than the MSE from ID2. Category III grouping yields smaller bias than grouping by ID1 in three out of four cases, the exception being CLIMP which forms less than half as many groups. Certain Category I variables yielded estimates with smaller MSE's. Clearly, random grouping should be avoided unless many groups can be formed and no Category II variable is readily available.

Three of four Category III variables yielded highly satisfactory estimates with small MSE's. The estimate from grouping on ACH2 is the most efficient of all estimates generated.

Only CLIMP among the Category III variables yielded an estimate with considerable bias and large MSE . The regression coefficients in Table 6.4 suggest that CLIMP acts as a suppressor when it enters the model with ACH and SRAA . As mentioned earlier, the small number

---

[5]In table 6.5, $\sqrt{MSE}$ was used instead of MSE for possible comparison with $SE(B_{\overline{YX}})$. In the discussion that follows $\sqrt{MSE}$ and MSE are interchangeable.

of groups formed by CLIMP also has detrimental effects on the precision of its estimate.

Three Category I variables, HSMATH, SAT2, and HSGPA2, yielded precise estimates of β_{YX} relative to the other Category I groupings. All have substantially larger zero-order correlations with ACH than with SRAA, and their between-groups standard deviations of ACH are large.

The remaining Category I variables, including SRAA2, yield estimates with large bias and large MSE. At the extreme (PARASP), $B_{\overline{YX}}$ is almost four times the ungrouped b_{YX}[6] and has a MSE 200 times the MSE of b_{YX}.

As Blalock and Hannan have stated, grouping on the dependent variable is disastrously biased. The unmeasured factors represented by the disturbance term in the initial linear model (Equation [3.1]) are confounded with the effects of the primary regressor to such a degree that the relation of ACH to SRAA is unrecognizable.

Fortunately, there are warning signals of poor estimation from Category I grouping, even when anonymously collected data prevent estimation of σ_{YX}. Of the eight Category I variables that yielded the largest biases, all except REPGPA had higher zero-order correlations with SRAA than with ACH (i.e., $r_{YZ} > r_{XZ}$). With SRAA2,

---

[6] We must re-emphasize that the superiority of a particular grouping variable is a function of the relation to be estimated. When we instead regress ACH on SRAA, for which $b_{YX} = .812$, grouping by ANTDEG ($B_{\overline{YX}} = .851$) and QCJOB ($B_{\overline{YX}} = .751$) result in small bias while grouping by HSPHYS ($B_{\overline{YX}} = 2.452$) and PARINC ($B_{\overline{YX}} = 1.848$) result in large bias. The standard errors for ANTDEG and QCJOB are also small for this regression. The question to be answered determines the quality of a particular characteristic for grouping purposes.

ANTDEG, QCJOB, or PARASP as the grouping variable, $\beta_{YZ \cdot X}$ even becomes larger than β_{XZ}. Also, the worst Category I variables (ANTDEG, QCJOB, PARASP) create a small $\hat{\sigma}_{\bar{X}}$, and distribute their observations unevenly among a few initial groups.[7]

2. Precision Independent of Scaling

There are no universal standards for judging what are acceptable values for bias and mean-squared error. The purposes for which an estimate is to be used determine what is "suitably precise". However, we can begin to set standards for acceptable bias and efficiency from grouping which are invariant under scalar transformations of variables.

We suggest that the investigator compare the predicted bias $(\hat{\theta})$ from a given grouping with twice the standard error of its corresponding estimate $(B_{\bar{Y}\bar{X}})$. If $\hat{\theta}$ is larger than 2 $SE(B_{\bar{Y}\bar{X}})$, drop the grouping variable from consideration. Selection among the remaining grouping variables can be based on the size of $\hat{\theta}$, on the efficiency of estimation, or on some other criteria (e.g., ease of collection or number of groups).

To judge the efficiency of a given estimate, the investigator can calculate $\widehat{Eff}(b_{YX}, B_{\bar{Y}\bar{X}}) = MSE(b_{YX})/MSE(B_{\bar{Y}\bar{X}})$. $\widehat{Eff}(b_{YX}, B_{\bar{Y}\bar{X}})$ should be as large as possible. Certainly, variables with efficiencies smaller than the worst of the Category IV variables should be excluded. As a further comparison, we suggest that the investigator calculated the ratio $\widehat{Eff}(b_{YX}, B_{\bar{Y}\bar{X}})/\widehat{Eff}(b_{YX}, B_{(m \text{ random groups of equal size})})$. This ratio provides some indication of the gain over random grouping in each case.

---

[7] The lowest two groups of ANTDEG's five groups contain fewer than 100 observations. Eighty-six (86) per cent of the observations on QCJOB are in two of its four categories. Ninety-seven (97) per cent of PARASP's were either "complete college" (4) or "obtain a graduate or professional degree" (5).

If we follow these guidelines in our example, we obtain the results depicted in Table 6.6. We have also compared the observed bias to $2\ SE(B_{\bar{Y}\bar{X}})$ in the table. With the $2\ SE(B_{\bar{Y}\bar{X}})$ as a criterion, all Category I variables are excluded, and all Category III and Category IV variables are retained, regardless of whether we look at observed bias or $\hat{\theta}$.

In every case, efficiency is small, but this occurs because of the small m values for almost every variable. If we compare the efficiency of each systematic grouping with that of grouping by ID1, we can exclude the worst Category III variable, which was previously retained. Furthermore, there are marked improvements in efficiency relative to random grouping for all Category III groupings and for the best of Category I grouping variables.

The variables that remain after applying exclusion principles for both bias and efficiency yielded estimates with the smallest biases and smallest MSE's overall. In Section 6.II.F, we suggest how the investigator might combine his best estimates when he does not wish to select among them.

E. Predicted Bias vs. Observed Bias.

Despite the specification and measurement errors, our predictions (Table 6.5) as to bias stood up well. For every grouping where the observed bias was greater than .2, $\hat{\theta}$ was greater than .2. With the exception of PARINC, the predicted bias was less than .1 whenever the observed bias was less than .1.

The prediction from $\hat{\theta}$ worked poorly only for ID1 and CLIMP. In the case of ID1, it is the sign reversal that troubles us and not the size of the error. There seems to be no reasonable explanation for the sign reversal other than the use of few groups with a random

Table 6.6. Comparison of estimates from grouped data using different criteria for acceptable bias from the regression of SRAA on ACH.

Grouping Variable (m)	|Observed Bias| <2 SE($B_{\cdot\cdot}$)	|Predicted Bias ($\hat{\theta}$)|[a] <2 SE($B_{\cdot\cdot}$)	$\widehat{Eff}(b_{YX}, B_{\cdot\cdot})$	$\frac{\widehat{Eff}(b_{YX}, B_{\cdot\cdot})}{\widehat{Eff}(b_{YX}, \text{random } Z_{(m)})}$
Category IV				
ID2 (100)	+	+	.040	1.08
ID1 (10)	+	+	.016	4.71
Category III[b]				
ACH2 (10)	+	+	.052	15.29
PARINC (10)	+	+	.024	7.06
HSPHYS (5)	+	+	.025	16.67
CLIMP (4)	+	+	.007	6.36
Category I[b]				
SRAA2 (5)	-	-	.002	1.33
HSMATH (5)	-	-	.027	18.00
SAT2 (13)	-	-	.020	4.44
HSGPA2 (23)	-	-	.018	2.20
POPED (6)	-	-	.008	4.21
REPGPA (7)	-	-	.008	3.64
NOBOOK (5)	-	-	.004	2.67
COLEFF (4)	-	-	.003	2.73
ANTIDEG (5)	-	-	.003	2.00
QCJOB (4)	-	-	.002	1.82
PARASP (5)	-	-	.002	1.33

[a] "+" = Within bounds of acceptable bias
"-" = Outside bounds of acceptable bias

[b] With the exception of ACH2 and SRAA2, variables within categories are ordered on the basis of observed bias (see Table 6.5).

grouping variable. In the previous section, we provided an explanation as to why estimates from CLIMP grouping might be disappointing (its suppressor relation with ACH and SRAA and its smaller number of groups); this explanation may also hold for poor prediction from CLIMP .

Every value of $\hat{\pi}$ proved to be larger than the observed bias. The Category III and Category IV variables along with the three Category I variables with the smallest bias yielded the lowest values of $\hat{\pi}$.

F. Composites of Estimates from Multiple Grouping Variables

The above findings suggest that an investigator can distinguish those grouping characteristics which lead to reasonably accurate estimates from those providing extremely misleading ones in empirical studies similar to ours. Once this separation has been accomplished, the investigator can choose a characteristic with small predicted bias. Better yet, he can use the available information about each characteristic and its expected bias to form a weighted composite of good grouped estimates. For example, he can weight grouped estimates in an inverse proportion to their predicted bias. Alternatively, he can give additional weight to the more stable estimates.

Table 6.7 provides two examples of composite extimates. In Example (A), we assume knowledge of σ_{YX} so that $\hat{\theta}$ can be used. In Example (B), σ_{YX} is treated as unknown, and thus the $\hat{\pi}$ values are used to weight the estimates. In each example, five of the seven grouping variables with the smallest predicted bias are used. We exclude ID2 as redundant with ID1 , and because it forms many more groups than any other variable. ACH2 is excluded on the grounds that compositing would be unnecessary if grouping on ACH2 were possible. Three sets of weights are determined: (1) in inverse proportion to the predicted bias, (2) inverse proportion to $SE(B_{\cdot\cdot YX})$ and (3) in inverse proportion to the

Table 6.7. Weighted Composites from grouped estimates of β_{YX} from the regression of SRAA on ACH.

Grouping Variable[a]	$B_{\bar{Y}\bar{X}}$[b]	$SE(B_{\bar{Y}\bar{X}})$	|Predicted Bias|	Weight (1)[c]	Weight (2)[d]	Weight (3)[e]
(A) Weights based on $\hat{\theta}$						
CLIMP	.717	.3971	.016	.243	.130	.162
ID1	.442	.1831	.075	.207	.195	.202
HSPHYS	.571	.0915	.095	.195	.222	.227
HSMATH	.414	.0248	.100	.192	.244	.222
PARINC	.558	.1314	.130	.174	.210	.188
Estimates yielded by the weights --				.562[f]	.531[f]	.538[f]
(B) Weights based on $\hat{\pi}$						
SAT2	.671	.0670	.210	.213	.229	.236
ID1	.442	.1831	.225	.211	.224	.226
PARINC	.558	.1314	.295	.199	.209	.202
HSMATH	.414	.0248	.307	.197	.242	.226
CLIMP	.717	.3971	.401	.180	.126	.111
Estimates yielded by the weights --				.568[f]	.566[f]	.554[f]

[a] In both examples, the grouping variables are ordered by the predicted bias ($\hat{\theta}$ or $\hat{\pi}$).

[b] The $B_{\bar{Y}\bar{X}}$ were transformed to Fisher Z's before weighting and averaging.

[c] Weight (1) = $\{[\Sigma(\text{Predicted bias } (Z_i))] - [\text{Predicted bias } (Z_i)]\} / 4\Sigma[(\text{Predicted bias } (Z_i)]$.

[d] Weight (2) = $\{\Sigma[SE(B_{\bar{Y}\bar{X}})_i] - SE(B_{\bar{Y}\bar{X}})_i\} / 4\Sigma[SE(B_{\bar{Y}\bar{X}})_i]$.

[e] Weight (3) = $\{[\text{Weight (1) for } Z_i][\text{Weight (2) for } Z_i]\} / [\Sigma(\text{numerator})]$.

[f] cf. $b_{YX} = .529$

predicted bias and $SE(B_{\overline{YX}})$. Since observations were initially standardized, the $B_{\overline{YX}}$ were transformed to Fisher Z's before weighting and averaging.

The resulting weighted composites are highly satisfactory. All composites were within .04 of b_{YX} . Estimate A(2) equals the estimates from grouping on the independent variable ACH2 . The remaining composite estimates do nearly as well, equaled or exceeded only by grouping on ACH2 , and in some cases, by grouping on ID2 and PARINC. Clearly, judicious weighting of grouped estimates can lead to precise estimation of the ungrouped regression coefficient.

III. Regression of Achievement on Aptitude

In our next example, we estimate the regression coefficient of achievement test performance (ACH) on aptitude test performance (SAT). Anonymity is not usually a problem in this case, but grouping could be economical. Thus we assume that σ_{YX} is known and limit discussion to the full-information situation.

This example will be considered in much less detail. Our primary purpose in this second empirical example is to illustrate that the suitability of a grouping variable depends on its relations with the main variables. We again standardized all variables prior to conducting the analysis.

A. Regression with Ungrouped Data

The equation relating ACH(Y) to SAT(X) is

$$ACH = .839(SAT)$$

with

$$SE(b_{YX}) = .0105 \quad ,$$

and

$$R^2_{Y \cdot X} = .704 \quad .$$

B. Categorization of Grouping Variables

Table 6.8 contains estimates for each grouping of the regression coefficients ($\hat{\beta}_{YX\cdot Z}$, $\hat{\beta}_{YZ\cdot X}$, $\hat{\beta}_{XZ}$, and $\hat{\beta}_{YZ}$) and their standard errors. The between-group standard deviation, $\hat{\sigma}_{\bar{X}}$, of SAT for the grouping variable is also included.

Again, we required an estimate of either $\hat{\beta}_{YZ\cdot X}$ or $\hat{\beta}_{XZ}$ to exceed three times its standard error to be considered significantly different from zero. The resulting categorization was as follows:

	$\hat{\beta}_{YZ\cdot X} \geq 3SE(\hat{\beta}_{YZ\cdot X})$		$\hat{\beta}_{YZ\cdot X} < 3SE(\hat{\beta}_{YZ\cdot X})$	
	Category I		Category III	
$\hat{\beta}_{XZ} \geq 3SE(\hat{\beta}_{XZ})$	HSGPA2	ANTDEG	SAT2	PARASP
	ACH2	HSMATH	PARINC	CLIMP
	REPGPA	HSPHYS	FATHED	QCJOB
	SRAA2	COLEFF	NOBOOK	
	Category II		Category IV	
$\hat{\beta}_{XZ} < 3SE(\hat{\beta}_{XZ})$	(NONE)		ID2	
			ID1	

Categories of several variables in the ACH-on-SAT regression differ from their categories with respect to the SRAA-on-ACH regression. ACH2, which now represents grouping on the dependent variable rather than on the independent variable, moves from Category III to Category I. HSPHYS also moves due to its correlation with ACH. The relative sizes of r_{YZ} and r_{XZ} again serve as useful clues to poor grouping variables since r_{YZ} is larger than r_{XZ} in six of the eight Category I groupings.

The number of variables in Category III is striking. Of the seven Category III variables in the ACH-on-SAT regression, five were in Category I in the regression of SRAA on ACH. The correlations of the Category III variables with ACH and SAT do not differ greatly

Table 6.8. Estimates of parameters relating SAT(X) and ACH(Y) to alternative grouping variables (Z)[a].

Variable Name	Group Size (m)	Parameter Estimates				
		$\hat{\beta}_{YX \cdot Z}$	$\hat{\beta}_{YZ \cdot X}$	$\hat{\beta}_{XZ}$	$\hat{\beta}_{YZ}$	$\hat{\sigma}_{\bar{X}}$
ID2	100	.839 (.0105)[b]	.014 (.0105)	.008 (.0193)	.020 (.0193)	.186
ID1	10	.839 (.0105)	-.003 (.0105)	-.046 (.0193)	-.042 (.0193)	.069
HSGPA2	23	.759 (.0116)	.164 (.0116)	.488 (.0169)	.535 (.0163)	.517
SAT2	13	.884 (.0662)	-.042 (.0662)	.987 (.0031)	.828 (.0109)	.989
ACH2	10	.082 (.0061)	.916 (.0061)	.827 (.0109)	.983 (.0035)	.835
PARINC	10	.838 (.0106)	.006 (.0106)	.076 (.0193)	.070 (.0193)	.146
REPGPA	7	.781 (.0117)	-.124 (.0117)	-.468 (.0171)	-.490 (.0169)	.498
POPED	6	.838 (.0106)	.007 (.0106)	.157 (.0191)	.139 (.0192)	.169
ANTDEG	5	.834 (.0106)	.039 (.0106)	.140 (.0192)	.156 (.0191)	.141
HSMATH	5	.765 (.0104)	.214 (.0104)	.346 (.0181)	.480 (.0170)	.349
HSPHYS	5	.811 (.0107)	.109 (.0107)	.257 (.0187)	.318 (.0183)	.294
NOBOOK	5	.844 (.0107)	-.025 (.0107)	.203 (.0189)	.146 (.0191)	.204
PARASP	5	.839 (.0106)	-.007 (.0106)	.087 (.0193)	.066 (.0193)	.101
SRAA2	5	.811 (.0123)	.054 (.0123)	.520 (.0165)	.476 (.0170)	.531
CLIMP	4	.838 (.0107)	.009 (.0107)	.165 (.0191)	.147 (.0191)	.185
COLEFF	4	.835 (.0106)	.039 (.0106)	.114 (.0192)	.134 (.0192)	.134
QCJOB	4	.838 (.0106)	.007 (.0106)	.118 (.0192)	.106 (.0192)	.123

[a] All variables have been standardized prior to grouping so that $\sigma_Y = \sigma_X = \sigma_Z = 1$, $\beta_{XZ} = \rho_{XZ}$, and $\beta_{YZ} = \rho_{YZ}$.

[b] Numbers in parenthesis are the standard errors of the regression coefficients.

in magnitude though the correlation of each Z with SAT is always larger than its correlation with ACH . The shift of SAT2 from Category I to Category III was expected; it now represents grouping on the independent variable. The remaining variables apparently enter Category III in part because of the strong correlation between ACH and SAT , which the model apportions to the independent variable SAT .

C. Estimates of Regressions from Different Grouping Methods

Table 6.9 contains estimated regression coefficients and other information. With a few minor exceptions, the results conform to our expectations.

The precision of Category IV grouping again is strongly related to the number of groups. The accuracy (bias) and stability (MSE) of grouping on ID2 is exceeded only by grouping on the independent variable (SAT2) . Grouping on ID1 yields a poorer estimate than grouping on ID2 , on any Category III variable, and on half of the Category I variables.

Category III grouping is clearly superior overall to grouping on variables from other categories. Observed bias is smaller than $2\ SE(B_{\bar{Y}\bar{X}})$ for 5 of 7 Category III variables. (See Table 6.10.) The exceptions are QCJOB and NOBOOK which form few groups with an uneven distribution of observations among the groups.

SRAA and COLEFF are the only Category I variables for which the observed bias falls within $2\ SE(B_{\bar{Y}\bar{X}})$. The estimates from the Category I variables, other than SRAA2 , in addition to yielding large bias, are about as inefficient as grouping on ID1 .

The decision rules discussed in Section 6.II.D are also useful with this example. If a variable is eliminated when (a) $|\hat{\theta}| \leq 2SE(B_{\bar{Y}\bar{X}})$ or (b) $Eff(B_Z;b) \leq Eff(B_{ID1};b)$, only NOBOOK among the Category III

Table 6.9. Estimates from grouped data of coefficients describing the regression of ACH on SAT.

Grouping Variable	Number of Groups (m)	$B_{\overline{YX}}$ [a]	Bias Observed	Bias Predicted from θ	$SE(B_{\overline{YX}})$ [a]	$\sqrt{\widehat{MSE}(B_{\overline{YX}})}$
CATEGORY IV						
ID2	100	.832	-.007	.003	.0590	.0594
ID1	10	1.053	.214	.029	.2168	.3036
CATEGORY III [b]						
SAT2	13	.838	-.001	-.001	.0190	.0190
PARINC	10	.817	-.022	.021	.0598	.0636
CLIMP	4	.876	.036	.042	.0388	.0528
POPED	6	.877	.039	.038	.0685	.0788
QCJOB	4	.912	.073	.054	.0216	.0775
PARASP	5	.744	-.095	-.059	.0903	.1310
NOBOOK	5	.718	-.121	-.174	.0372	.1266
CATEGORY I [b]						
ACH2	10	1.168	.329	.329	.0541	.3338
SRAA2	5	.899	.060	.072	.0543	.0809
REPGPA	7	1.019	.180	.176	.0418	.1848
COLEFF	4	1.054	.213	.241	.1169	.2438
HSGPA2	23	1.057	.218	.219	.0329	.2205
ANTDEG	5	1.120	.281	.271	.0607	.2875
HSPHYS	5	1.237	.398	.296	.0422	.4002
HSMATH	5	1.396	.557	.531	.0478	.5590

[a] Estimates from ungrouped data: $b_{YX} = .839$; $SE(b_{YX}) = .0105$.

[b] With the exception of ACH2 and SAT2, variables within categories are ordered on the basis of observed bias.

Table 6.10. Comparison of estimates from grouped data using different criteria for acceptable bias in the regression of ACH on SAT.

Grouping Variable (m)	$\lvert$Observed Bias$\rvert$ ≤ 2 SE(B_{--})	$\lvert$Predicted Bias (θ)$\rvert$[a] $\leq$ SE(B_{--})	$\widehat{Eff}(b_{YX}, B_{--})$	$\frac{\widehat{Eff}(b_{YX}, B_{--})}{\widehat{EFF}(b_{YX}, \text{random } Z_{(m)})}$
Category IV				
ID2 (100)			.177	4.78
ID1 (10)	+	+	.034	10.00
Category III[b]				
SAT2 (13)	+	+	.553	122.89
PARINC (10)	+	+	.165	48.53
CLIMP (4)	+	+	.198	180.00
POPED (6)	+	+	.133	70.00
QCJOB (4)	-	+	.135	122.73
PARASP (5)	+	+	.080	53.33
NOBOOK (5)	-	-	.083	55.33
Category I[b]				
ACH2 (10)	-	-	.031	9.12
SRAA2 (5)	+	+	.130	86.67
REPGPA (7)	-	-	.057	25.91
COLEFF (4)	+	-	.043	28.67
HSGPA2 (23)	-	-	.048	5.85
ANTDEG (5)	-	-	.037	24.67
HSPHYS (5)	-	-	.026	17.33
HSMATH (5)	-	-	.019	12.67

[a] "+" = Within bounds of acceptable bias.

"-" = Outside bounds of acceptable bias.

[b] With the exception of ACH2 and SAT2, variables within categories are ordered on the basis of observed bias. (See table 6.9).

variables is eliminated and all the Category I variables except SRAA2 are dropped.

D. Predicted Bias vs. Observed Bias

The results of the predictions in the regression of ACH on SAT are as satisfactory as the results in the earlier example. The prediction from ID1 grouping is again among the most errant. In general, however, grouping characteristics which produce good estimates can be selected on the basis of predicted bias, especially when the standard errors of the grouped estimates are also taken into account.

IV. Summary of Empirical Results

We set out in Chapter 6 to demonstrate the utility of the grouping concepts and methods developed in Chapters 3 and 4 under realistic empirical conditions. The empirical evidence regarding the estimation of β_{YX} conformed to the predictions from the principle of incorporating the grouping characteristics as variables in the structural model, which, in turn, lead to the taxonomic categorization of grouping variables. The latter classification resulted in clusters of readily identifiable "good" and "bad" grouping variables under most aggregated conditions. We further showed that if the investigator formed a weighted composite of estimates from several of his best grouping variables, his resulting estimate is invariably highly accurate.

Thus we demonstrated some effective strategies of estimating simple linear regression coefficients (and zero-order correlation coefficients) when data aggregation is under the investigator's control and the grouping characteristics under consideration have at least an interval scale. To a certain degree, our results are generalizable to naturally aggregated data where some degree of disaggregation is

feasible. The possibilities of utilizing nominally scaled grouping characteristics were discussed in Chapter IV, but the procedures suggested for such variables were not demonstrated empirically.

CHAPTER 7

SUMMARY AND CONCLUSIONS

I. Summary of Findings

We have examined certain consequences of estimating regression coefficients at the level of individuals from aggregated data. In Chapter 1, various research contexts in which such questions arise were described and the main emphasis of our investigation was identified.

In Chapter 2, we reviewed previous literature on grouping in the two-variable case. The literature on estimating both correlation coefficients and regression coefficients was considered.

In Chapter 3, we discussed the various factors which affect the estimation of the simple linear regression coefficient and zero-order correlation coefficient when data are grouped on some interval variable. With one exception, it was assumed throughout that there were no measurement errors in X. Though speaking in terms of "structural equation models" is somewhat awkward when there are only two variables involved, this term was used because the bivariate regression was simply a special case of a multivariate structural model.

We first demonstrated that the estimate of β_{YX} ($B_{\overline{YX}}$) from grouped data is unbiased if the assumptions regarding the disturbances in the simple model used by earlier investigators are satisfied. However, the slope estimates from grouped data were shown to be less efficient than the estimates from ungrouped data. This finding led to the criterion of maximization of the between-groups variance (minimization of the within-group variance) of the independent variable as an appropriate method of judging the efficiency of alternative grouping procedures.

The investigation was then expanded to consider in greater detail the concept of grouping by a "grouping variable". This logic suggested that the criterion by which the individual observations are to be grouped can be treated as a random variable which may be related to other variables in the structural equation system. Furthermore, the system specified that the grouping variable Z, if related to another variable, is prior to that variable. The alternative relations of the grouping variable to the dependent and independent variables were then used to generate a four-category taxonomy which included all grouping variables satisfying a specific set of relational restrictions imposed by that category.

The estimates from data grouped by Category I variables (Z related to both $Y \cdot X$ and X) were found to biased. This apparent disagreement between the simple model and our alternative structure can be explained by the misspecification of the simple model when the grouping variable is directly related to both dependent and independent variables. Further examination of this phenomenon led to a recommendation that the relation between the grouping variable and the dependent variable be minimized. Grouping on Category I or Category II variables was discouraged because such variables are directly related to $Y \cdot X$, and few variables can be expected to meet the necessary criteria that Z be unrelated to X and $\Sigma \bar{X}^2$ be nonzero at the same time.

The relative efficiencies of variables from the different categories were also examined. It was determined that Category III grouping variables (Z related to X but not to $Y \cdot X$) yield the most efficient grouping procedures so long as there are variables with efficiencies greater than Category IV variables [whose efficiencies are on the order of $(m-1)/(N-1)$ where m is the number of groups and N, the total

number of observations.]. It was suggested that for certain values of $\beta_{YZ \cdot X}$ and $\beta_{.Z}$, Category I grouping, though slightly biased, can yield more efficient estimates than either Category II or Category IV grouping.

We examined the possible causes of variations in the magnitude of bias and the relative efficiency within categories of the taxonomy and the special problems in grouping by nominal characteristics in Chapter 4. The within-variable properties considered were (1) the coarseness of grouping, (2) the distribution of observations among the groups, and (3) the distribution of the independent variable within and among the groups. As might be expected, the most efficient estimates were found to coincide with variables that generated a large number of maximally discrete and compact groups.

We also considered ways of applying "structural equation" methods with nominal grouping characteristics in Chapter 4. A classification scheme proposed by Wiley was discussed wherein grouping variables are categorized by their scale (nominal or interval) and by whether the groups in the study are the entire population (fixed) or only a sample from the population of interest (random). The nominal grouping variable Z^+ was viewed as a surrogate for an underlying grouping variable $Z^=$ has a metric. Though $Z^=$ is latent and unmeasurable, it can be estimated by classification procedures describing group differences in Z^+. Sampling bias was said to affect grouped estimates when the classes of the grouping variable are unrepresentative of the population.

Dummy coding procedures used by economists were suggested as a way to incorporate the nominal grouping characteristic in our models. Dummy coding is less time-consuming and complex than Wiley's procedure. It yields functions which can be compared directly with the parameters generated by ordered grouping characteristics.

In Chapter 5, we described various procedures for analyzing the effects of grouping in the multivariate case. Of particular interest was a statistic developed by Feige and Watts (1972) for assessing the divergence between grouped and ungrouped regression coefficients. Also, we showed that the results from the "structural equation" approach in the two-regressor case agreed with the findings in the single-regressor case. Extension of the results from the "structural equations" approach to more than two regressors is straightforward. However, the analysis rapidly becomes complicated with additional regressors because of the necessity to specify the structural relations among all variables in the model (including the grouping variable).

Empirical examples of grouping in the single-regressor case were presented in Chapter 6. In general, the results conformed to our expectations and the predictions from the structural equations approach were reasonably accurate. The use of weighted composites of estimates from different grouping methods was demonstrated. These weighted composites were recommended as a possible means of estimating coefficients when information on certain primary variables is collected anonymously.

When the within-category effects of the different factors are combined with our knowledge of the category and scale differences, several principles evolve for selecting a grouping variable which minimizes bias and maximizes efficiency. A partial list of these principles in the single-regressor case includes the following:

A. To obtain unbiased estimates of the linear regression coefficient, choose a Z so that (in order of preference)

   1) Z is related to X but not to $Y \cdot X$ (Category III),

   2) Z is not related to either X or Y (Category IV),

or 3) Z is related to $Y \cdot X$ but not to X (Category II).

Category III variables are preferable because they yield generally efficient estimators because the between-group variation in the regressor is maximized.

B. When biased estimates are the only alternative, choose Z so that

1) β_{XZ} is as large as possible,

2) $\beta_{YZ \cdot X}$ is as small as possible,

3) β_{YZ} is smaller than β_{XZ}, and

4) the ratio $\sigma_{\bar{Z}}/\sigma_{\bar{X}}$ approaches as near as possible the ratio σ_Z/σ_X.

C. The efficiency of the grouped estimator increases as

1) m approaches N; or

2) average n increases when random measurement errors in X are possible, but decreases otherwise; or

3) the correlation ratio η_X^2 approaches unity; or

4) the pooled within-group variance in the independent variable becomes smaller; or

5) the degree of overlap among the within-group distributions of the independent variable decreases.

There are obviously other intangibles that cannot be dealt with by general principles. There is always the problem of degree of investigator control over the grouping process. As stated earlier, anonymous collection of data on some primary variables seriously complicates matters as does adding more regressors. We have tried to identify only the strategic aspects of the process of determining the effects of grouping and have left to future investigations the practical details of application. Proper application of these principles requires that the investigator thoroughly understand the theoretical model in

question, and no set of guidelines can adequately ensure that this will occur.

II. Suggestions for Further Investigation

At a number of points, we have noted areas where the present state of knowledge on the complications due to data aggregation is weak and further investigation is warranted. Here we indicate several of the more interesting and pressing questions.

1. Nominal Grouping Characteristics -- In the introductory chapter, we described five research problems in which aspects of data aggregation are encountered. The discussion that followed focused almost entirely on questions that arise in two contexts [(C) economy of analysis and (D) anonymously collected data]. Perhaps the most important question from the perspective of educational researchers is how to determine the effects of grouping on a nominal characteristic such as school [problem (E)]. Our treatment of nominal variables in Chapter 4 merely provides some suggestions about how this work might proceed. Much more research is necessary to determine the special complications that arise in predicting the effects of grouping on a nominal characteristic.
2. Missing Data and Measurement Errors -- The suggested utilization of grouping in handling problems with missing data [problem (A)] and measurement errors [problem (B)] requires further elaboration and investigation.
3. Weighted Composites and Anonymously Collected Data -- The description of the use of aggregated data to overcome complications with anonymously collected data [problem (D)] and the

subsequent example which used weighted composites of estimates from grouped data to estimate coefficients represent a potentially valuable new field for planned application of data aggregation. More work is necessary to establish the generality of the technique of estimating individual-level relations from weighted composites of between-group coefficients.

4. Multivariate Models -- A more thorough investigation of the effects of grouping in models with multiple regressors is highly desirable. The comparative utility of the "structural equations" approach and the procedures suggested by Feige and Watts needs to be investigated. Additionally, hardly anything is known about the optimal grouping method when the hypotheses of interest posit some form of simultaneity of causation in multivariate models.

5. Aggregation Over Time -- An investigation of whether principles developed here apply when the grouping variable is some time interval ("year", "occasion") would be of value. The results may provide new insight into the partitioning of observation periods in classroom process studies.

6. Appropriate Model Specification -- We have purposely focused on the conditions under which the estimates from grouped data provide accurate or misleading information about relations among measurements on individuals. It is evident that the principles governing aggregation bias are a subset of the problems that appear in econometrics literature under the heading of "specification bias". The necessary interrelation between specification bias and aggregation bias needs to be elaborated and communicated to the educational research

community. This elaboration would necessarily include warnings about the potential hazards of accepting global measures of association (e.g., individual-level correlations) as accurately reflecting the actual processes in operation. When there exist group-to-group differences on the primary variables, it is often more appropriate to conduct within-group analyses or to include additional variables that account for group differences in the model. This latter kind of a specification problem suggests the interface between the analysis of covariance procedures and the analysis of grouping effects.

7. Multilevel Analysis -- In the literature on school effects, investigators have begun to recognize that it may be necessary to adjust for the lack of independence among students within classrooms. Procedures that combine within-class or within-school analyses with analyses at a higher level of aggregation deserve more attention.

This list is by no means complete. However, it does accurately reflect the concerns over data aggregation in educational research and directions for further inquiry by educational researchers.

BIBLIOGRAPHY

Alker, H.R. Jr. "Typology of Ecological Fallacies." In M. Dogan and S. Rokkan (Eds.), Quantitative Ecological Analysis in the Social Sciences. Cambridge, Mass.: MIT Press, 1969, pp. 69-86.

Allardt, E. "Aggregate Analysis: The Problem of Its Information Value." In M. Dogan and S. Rokkan (Eds.), Quantitative Ecological Analysis in the Social Sciences. Cambridge, Mass.: MIT Press, 1969, pp. 41-51.

Averch, H.A., Carroll, S.J., Donaldson, J.S., Kiesling, H.S., and Pincus, J. How Effective is Schooling? A Critical Review and Synthesis of Research Findings. Santa Monica, CA.: The Rand Corporation, 1972, R-956-PCS/RC.

Bartlett, M.S. "Fitting a Straight Line When Both Variables are Subject to Error." Biometrics, 1949, 5: 207-212.

Blalock, H.M. Causal Inferences in Nonexperimental Research. Chapel Hill, NC.: University of North Carolina Press, 1964.

_____, Wells, C.S., and Carter L.F. "Statistical Estimation with Random Measurement Error." In E.F. Borgatta and G. Bohrenstedt (Eds.), Sociological Methodology 1970. San Francisco, CA.: Jossey-Bass, 1970, pp. 75-103.

Boruch, R.F. "Maintaining Confidentiality of Data in Educational Research: A Systematic Analysis." American Psychologist, 1971, 26: 413-430.

Burstein, L. "Issues Concerning Inferences from Grouped Observations." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Ill., April, 1974.

_____. "Data Aggregation in Educational Research: Applications." Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C., April, 1975.

Cartwright, D.S. "Ecological Variables." In E.F. Borgatta (Ed.), Sociological Methodology 1970. San Francisco, CA.: Jossey-Bass, 1970, pp. 155-218.

Cramer, J.S. "Efficient Grouping: Regression and Correlation in Engle Curve Analysis." Journal of the American Statistical Association, 1964, 59: 233-250.

Duncan, O.D., Cuzzort, R.P., and Duncan, B.D. Statistical Geography. Glencoe, Ill.: Free Press, 1961.

_____ and Davis, B. "An Alternative to Ecological Correlation." American Sociological Review, 1953, 18: 665-666.

Elashoff, J.D. and Elashoff, R.M. "Missing Data Problems for Two Samples on a Dichotomous Variable." Stanford, CA.: Stanford Center for Research and Development, 1971, Memorandum No. 73.

Estes, W.K. "The Problem of Inferences from Curves Based on Grouped Data." Psychological Bulletin, 1956, 53: 134-140.

Feige, E.L. and Watts, H.W. "Protection of Privacy Through Microaggregation." In R.L. Bixco (Ed.), Data Bases, Computers and the Social Sciences. New York, NY.: John Wiley & Sons, Inc., 1970, pp. 261-272.

_____. "An Investigation of the Consequences of Partial Aggregation of Micro-Economics Data." Econometrica, 1972, 40: 343-360.

Gehkle, C. and Biehl, R. "Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material." Journal of the American Statistical Association Supplement, 1934, 29: 169-170.

Goldberger, A.S. Econometric Theory. New York, NY.: John Wiley & Sons, Inc., 1964.

_____. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." Madison, Wisc.: The University of Wisconsin, Institute for Research on Poverty, 1972, Discussion Paper No. 123-72.

Hannan, M.T. "Problems of Aggregation and Disaggregation in Sociological Research." Chapel Hill, NC.: University of North Carolina, Institute for Research in the Social Sciences, 1971, Methodology Working Paper No. 4.

_____. Aggregation and Disaggregation in Sociology. Lexington, Mass.: Heath, 1971.

_____. "Approaches to the Aggregation Problem." Stanford, CA.: Stanford University, Laboratory for Social Research, 1972, Technical Report No. 46.

_____ and Burstein L. "Estimation from Grouped Observations." American Sociological Review, 1974, 39: 374-392.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. Sample Survey Methods and Theory, Volume II: Theory. New York, NY.: John Wiley & Sons, Inc., 1953.

Iversen, G.R. "Recovering Individual Data in the Presence of Group and Individual Effects." American Journal of Sociology, 1973, 79: 420-434.

Johnston, J. Econometric Methods. New York, NY.: McGraw-Hill, 1972, 2nd Ed.

Kline, G.F., Kent, K., and Davis, D. "Problems in the Causal Analysis of Aggregate Data with Applications to Political Instability." In J. Gillespie and B. Nesvold (Eds.), Macro-Quantitative Analysis. Beverly Hills, CA.: Sage Publications, 1971, pp. 251-279.

Mandansky, A. "The Fitting of Straight Lines when Both Variables are Subject to Error." American Statistical Association Journal, 1959, 54: 173-205.

Menzel, H. "Comment on Robinson's 'Ecological Correlation and the Behavior of Individuals'." American Sociological Review, 1950, 15: 674.

Prais, S.J. and Atichinson, J. "The Grouping of Observations in Regression Analysis." Review of the International Statistical Institute, 1954, 22: 1-22.

Riley, M.W. "Sources and Types of Sociological Data." In R.L. Faris (Ed.), Handbook of Modern Sociology. Chicago, Ill.: Rand McNally, 1964, pp. 1014-1020.

Robinson, W.S. "Ecological Correlations and the Behavior of Individuals." American Sociological Review, 1950, 15: 351-357.

Scheuch, E.K. "Cross-National Comparisons Using Aggregate Data: Some Substantive and Methodological Problems." In R.L. Merritt and S. Rokkan (Eds.), Comparing Nations: The Quantitative Data in Cross-National Research. New Haven, Conn.: Yale University Press, 1966, pp. 148-156.

Selvin, H.C. "Durkheim's Suicide and Problems of Empirical Research." American Journal of Sociology, 1958, 63: 607-619.

Shively, W.P. "'Ecological' Inference: The Use of Aggregate Data to Study Individuals." American Political Science Review, 1969, 63: 1183-1196.

Theil, H. Linear Aggregation in Economic Relations. Amsterdam: Holland Publishing Company, 1954.

Thorndike, E.L. "On the Fallacy of Imputing the Correlations Found for Groups to the Individuals and or Smaller Groups Composing Them." American Journal of Psychology, 1939, 52: 122-124.

Wald, A. "Fitting of Straight Lines if Both Variables are Subject to Error." Annals of Mathematical Statistics, 1940, 11: 284-300.

Walker, H.M. "A Note on the Correlation of Averages." Journal of Educational Psychology, 1928, 19: 636-642.

Werts, C.E. and Linn, R.L. "Considerations When Making Inferences Within the Analysis of Covariance Model." Educational and Psychological Measurement, 1971, 31: 407-416.

Yule, G.U. and Kendall, M.G. An Introduction to the Theory of Statistics. London: Griffin, 1950, 14th Ed.